UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

CORSO DI LAUREA MAGISTRALE IN ECONOMICS AND FINANCE

TESI DI LAUREA

MACHINE LEARNING APPLICATIONS TO ECONOMICS: PREDICTIONS AND HETEROGENEOUS CAUSAL EFFECTS

RELATORE:

CH.MO PROF. LUCA NUNZIATA

LAUREANDO: MARCO MUSUMECI

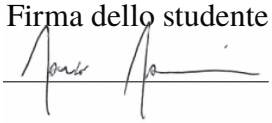
MATRICOLA N. 1150319

ANNO ACCADEMICO 2018 – 2019

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere. Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell'elaborato sono stati indicati nel testo e nella sezione "Riferimenti bibliografici" e che le eventuali citazioni testuali sono individuabili attraverso l'esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente



Abstract

This thesis presents two applications of machine learning techniques in the field of economics. In the first chapter, we review from a theoretical point of view the machine learning models implemented in this work, i.e. LASSO and tree-based methods. In the second chapter, we study how the predictive power of machine learning models can be exploited to efficiently target smoking prevention policies. Finally, in the third chapter, we adopt the so-called causal forest approach to explore heterogeneity in treatment effects in the STAR Tennessee Project, that aims at assessing the impact of belonging to small classes on students' performance in standardized tests.

Contents

Introduction

The use of machine learning (ML) algorithms is becoming more and more frequent in a variety of research fields, from medicine to finance. Economics is no exception.

The reasons for this growth are various. First, today it is easier than ever to get access to data with billions of observations or millions of features. These data need new tools to be analyzed and machine learning is one of the most effective. Secondly, these algorithms provide a powerful and flexible way to produce quality predictions. Often these forecasts turn out to be more accurate than the ones produced via traditional econometric models. Finally, the technological progress has made available in a few minutes computations that would have taken hours only ten years ago. All these factors contributed to an exponential rise in the use of machine learning in the economic literature.

In this thesis we review some of the most well-known machine learning algorithms (LASSO and tree-based methods) and study different applications in the area of economics. The most straightforward way to implement ML in this field is to exploit its strength: improve predictive power. Even if economics is usually focused on causal inference, we analyse different frameworks in which accurate forecasts turn out to be more useful than inference (Kleinberg et al. 2015). In particular, we make use of the British Cohort Study of 1970 to predict the group of individuals who have the highest probability to start smoking between the ages of 10 and 16. Being able to identify this group, makes it possible to target appropriately a mentoring intervention to increase awareness of the risks connected to smoking.

Even if predictions are the most direct way to use ML, a growing literature is adapting it to the purpose of producing valid causal inference. Specifically, we focus on the so-called causal forest approach with the aim of studying heterogeneous causal effects (HTE), as proposed in Wager and Athey 2017. This method is adopted to study HTE in the STAR Tennessee Project. This project aimed at evaluating the effect of belonging to a class of small size on the students' performance on standardized tests. Thanks to causal forest, we will be able to analyze the heterogeneity of treatment effects in a deeper way than previous studies.

The organization of this work is the following: in the first chapter we present the machine

learning tools implemented in this thesis and the strategy adopted to evaluate the quality of the predictions produced. In the second chapter we propose our application on the British Cohort Study that aims at predicting adolescents' smoking status. Finally, in the third chapter, we implement causal forest to study heterogeneous treatment effects in the STAR Tennessee Project.

Chapter 1

Theoretical background

1.1 The use of machine learning in economics

The strength of machine learning is that it can discover complex patterns that drive the behavior of a certain variable of interest. Instead of concentrating on causal inference, machine learning algorithms are build to maximize the predictive power of a model. As Mullainathan and Spiess 2017 point out, even when these algorithms produce regression coefficients, no causal interpretation can be ascribed to them. Therefore, as the focus of most economics applications is on inference, we need to understand how machine learning can be applied in this field.

1.1.1 Prediction policy problem

The most straightforward use of ML is in the context of policy problems, where in many cases producing accurate forecasts can play a central role in the empirical research.
First of all, we need to understand what is the difference between a policy application that requires inference and another that requires predictions. The best way of doing it is with an example. Let us consider a policy maker that must decide whether to continue a program that aims at preventing high-risk young individuals to commit crimes. In this case what we want to know is if those who attended the program are less likely to commit crimes, as a consequence of attending it. This is clearly a causal inference problem.
Consider now a strictly connected issue: a policy maker must decide which individuals should be involved in this program. In other words, he should understand which are the individuals with the highest risk of committing a crime, because they are the ones to whom we want to target an intervention. Here we are not interested in uncovering the causal drivers of criminality, we just want to predict who is likely to be involved in an illegal action (a similar problem is presented in

the paper by Chandler, Levitt, and List 2011[1], even if the authors do not use machine learning). This second example is a prediction problem.

Most of the literature on policy application is focused on causal inference, but as we have seen, predictions should also play an important role in the decision-making process when implementing a policy. As shown by Mullainathan and Spiess 2017 the statistical tool that are used to make inference are often outperformed by machine learning methods when the focus is on predictive power. Therefore, in some cases the use of ML could lead to policies that are better designed. It is worth stressing that inference and predictions should not be considered two contrasting approaches. Instead, they should be seen as two complementary tools that are used to answer different questions. Kleinberg et al. 2015 underlines these points. The authors show that prediction problems are extremely common and that they can involve a wide range of fields such as education, labor market and social policies or regulation. Moreover, they provide an example of prediction policy problem implementation in the field of health economics, that we briefly outline here. They show that a machine learning model could help in deciding on which people should undergo a certain surgery to cure osteoarthritis, based on their life expectancy after the surgery. Given that the surgery considered needs numerous months before it can actually improve the quality of life of the patient, it is clear that, if an individual has a lower probability of living long enough to take benefit from it, he would be better off by avoiding it. The authors build a prediction model that determines who should not undergo the surgery because of a high risk of dying in the subsequent months. The group that they identify as the 1% most at risk of dying within 12 months from the surgery presents a mortality rate of 56% against an average of the sample of 4.2%. Excluding from the surgery those that are in the 1% with the highest risk would first of all improve their quality of life, but also save economic resources.

This example should clarify that predictions are sometimes more appropriate than causal inference. Indeed, we were not interested in determining the causal relation between certain characteristics of an individual and its response to the surgery, we were just interested in forecasting his level of risk.

1.1.2 Causal inference

Even if predictions are the most direct way to use ML in economics, a recent and growing literature is developing different approaches to adapt machine learning techniques for statistical inference. A first approach is the one proposed by McCaffrey, Ridgeway, and Morral 2004

[1]The authors try to identify a group of Chicago Public Schools students with a high probability of being involved in shootings. This is clearly a prediction problem, because we don't need to uncover a causal relationship between being involved in shootings and certain individual characteristic, but we only need to find those individuals that are at high risk.

where they develop a propensity score matching model using ML. Even if this method performs well in different simulated data studies, it has been criticized by Athey and G. W. Imbens 2017 because it can produce biased estimates in many cases.

Another model is the one proposed in Belloni, Chernozhukov, and Hansen 2014, where the authors develop a machine learning approach, based on LASSO, to check for the robustness of a model to an omitted variable bias. They implement their methodology to the well-known paper by Acemoglu, Johnson, and Robinson 2001, that studies the effect of institutions on the levels of aggregate output. They show that the original model specification is not robust to the implementation of the controls selected via machine learning.

Finally, a quite large part of the literature focused on ML is concerned with the estimation of heterogeneous treatment effects (HTE). Various methods have been developed to study this topic. We will concentrate in this thesis mainly on the approach proposed by Athey, Imbens and Wager in different papers[2]. Their focus is on the estimation of HTE using CARTs and random forest.

1.2 Machine learning models

In this section we review the machine learning instruments that have been used in the development of this thesis. Subsequently, we look at the methods applied to evaluate the performance of a model in terms of accuracy.

This review does not have the objective of presenting formally the ML technique used but it aims at giving an overview on these concepts and showing with examples how they can be applied. A formal explanation of how all these techniques work can be found in Hastie, R. Tibshirani, and Friedman 2009[3]. The notation used in this chapter is the one proposed by James et al. 2013.

1.2.1 LASSO

The least absolute shrinkage and selection operator (LASSO), introduced by R. Tibshirani 1996, is probably the most well-known machine learning tool in Economics. The LASSO is extremely similar to a least squares regression. Its coefficients are estimated minimizing the same quantity of OLS plus a penalization term that grows proportionally to the increase of the absolute values sum of the coefficients estimated. In formula LASSO estimates are found minimizing the quantity in equation 1.1.

[2]The most relevant papers for the study of their model are Athey and G. Imbens 2016, Wager and Athey 2017 and Athey, J. Tibshirani, and Wager 2016.

[3]A less technical compendium is represented by James et al. 2013, while a precious resource for the implementation on the software R (most of the presented techniques are not available in Stata) is Lantz 2015.

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{1.1}$$

The advantage of adding a penalization term is that, depending on an appropriately selected parameter λ, it can shrink to zero the estimates of some of the coefficients. As pointed out by the author, this characteristic of LASSO has two advantages: it can improve the prediction quality in a hold-out sample reducing overfitting, and it can help interpretation, by selecting a limited number of characteristics that can predict the outcome. This last point is particularly useful when we are working with high dimensional data (i.e. data where the number of features is higher than the number of observations), where traditional econometric techniques cannot produce estimates.

The addition of a penalization term also introduces the question of how to determine it. This problem occurs often in the field of machine learning and it is usually solved choosing its value by cross validation. It is important to remark that if λ is equal to zero, then we are simply performing an OLS regression.

In order to provide a clearer picture of how LASSO works, when it is particularly useful, and how to determine λ, we propose an example with real word data. This example consists of a dataset containing the information on students' performance in a final year test of a Portuguese school and 25 variables that describes different characteristics of the students [4]. The number of individuals considered is 349 and our goal is to predict the outcome of the final test, based on the students' characteristics. Hereafter, we will refer to this data as the Portuguese school data. First of all, we need to randomly select one half of the observations as a training sample so that it will be possible to evaluate the performance of the model in a hold-out sample. Therefore, in our example we have a training sample of 174 observations and a test sample of 175. Secondly, we need to determine the value of λ. The importance of this last point is underlined by figure 1.1, that shows how the coefficients associated to each of the 25 features in the data varies as λ moves from zero to two.

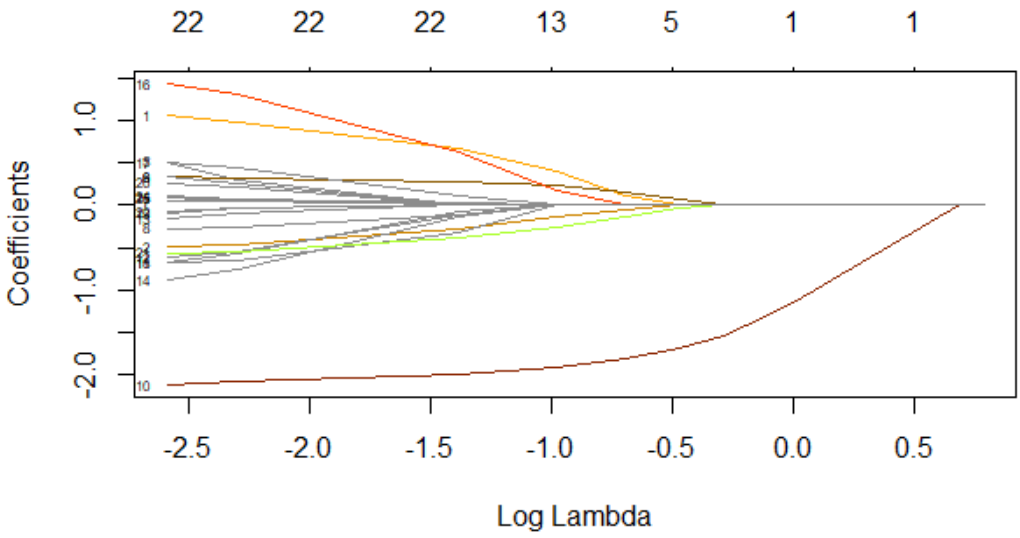
Figure 1.1 shows clearly that choosing an appropriate level of λ is crucial. When λ is zero we are running an OLS regression. However, as λ increases, more and more coefficients are shrank to zero, up to a point where all of them are equal to zero. It is interesting to notice that for a certain value of λ most of the coefficients (the ones colored in gray) are reduced to zero and only six of them (the ones colored differently) remain positive. These six characteristics selected are students' age and sex, father and mother's education, the number of previous class failed

[4] This dataset is a subsample of the one provided by Cortez and Silva 2008.

and a variable that captures the frequency of going out with friends. We can expect these six characteristics to have together a good predictive power of the final exam grade that a student will obtain.

Figure 1.1 is useful to understand how coefficients are affected by λ, but it does not say anything about which is the optimal level of this parameter that we should select if our aim is to maximize the precision of the predictions produced. The best way to achieve this goal is by cross validation[5]. Therefore, we determined the optimal value of lambda implementing a 10-fold cross validation.

Figure 1.1: LASSO coefficients as a function of λ



The graph shows how the values of coefficients (represented in the y-axis) changes as the value of lambda (bottom x-axis) goes from 0 to 2. The top x-axis represents the number of coefficients different from zero.

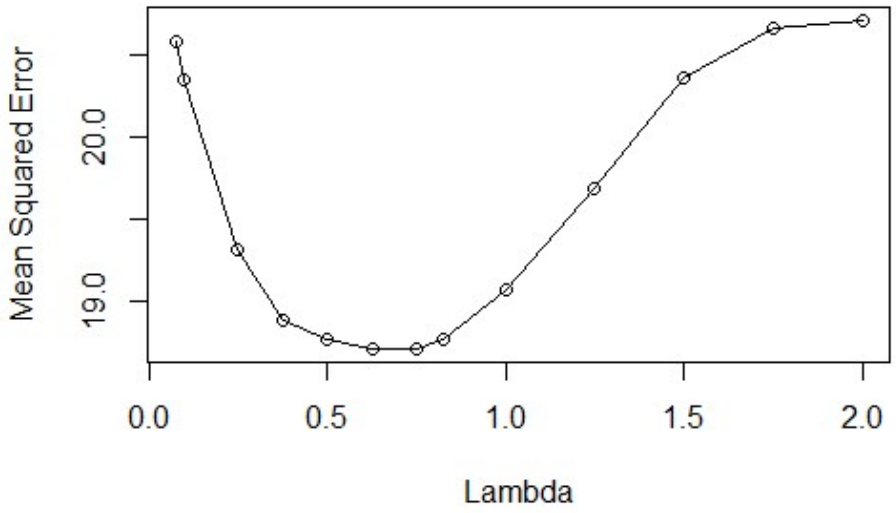
Figure 1.2 describes how in our example the cross validated Mean Squared Errors (CV MSE) varies as lambda goes from zero to two.

The graph has a U-shape, meaning that initially the variable selection increases cross validated predictions quality, but after a certain point it starts to decrease it (this fact is expected because with the highest value of lambda all the coefficients tend to zero). The optimal value of lambda selected by cross validation is 0.625 (i.e. the one that minimize CV MSE). This value is the one that should maximize out-of-sample predictive power of the LASSO regression. Consequently, it is now used to run the LASSO regression on the Portuguese school dataset training sample.

[5]We are considering a ten-fold cross validation. For more details on how cross validation works and the different type of cross validations that can be implemented see chapter 5 of James et al. 2013.

The results of this regression are the following: when implementing LASSO with $\lambda = 0.625$, only three coefficients are different from zero (mother's education, frequency of going out with friends and number of previous class failed). This means that, in order to obtain the best out-of-sample predictive power, it seems to be necessary a strong variable selection in this dataset. In other words, we expect that the best way of predicting the final grade of a student is by using only the three above-mentioned variables instead of the entire set of characteristics that the dataset contains.

Figure 1.2: CV MSE as a function of lambda



The graph shows how the CV MSE vary as the value of lambda changes from 0 to 2 in the Portuguese school dataset.

A final remark concerns the possibility to deal with non-linearities and interaction terms with LASSO. As high dimensionality does not cause overfitting, we can include in the model second or even third grade polynomial of the features considered. Moreover, it is possible to include all the interaction terms between the variables of the model. Thus, the number of variables considered will increase in a way that causes traditional econometrics tools not to work properly. However, LASSO can select only those variables with a strong predictive power of the dependent variable. Therefore, if we believe that interaction terms may play an important role in determining the outcome, LASSO turns out to be a powerful instrument because we can include them in the model without worrying on the consequences of the resulting high dimensionality.

The final step of this analysis is to test the predictive power of the model(s) implemented using the test sample. Even if cross validated MSE is often a good approximation of the performance on never-seen-before data, the best way to evaluate the quality of a model predictions is with

out-of-sample data. This analysis is developed later in the chapter, when we compare the performance of the different models implemented and explain how to do it.

1.2.2 Classification and Regression Trees

We now review the so-called Classification and Regression Trees (CARTs) that were introduced by Breiman 2017.

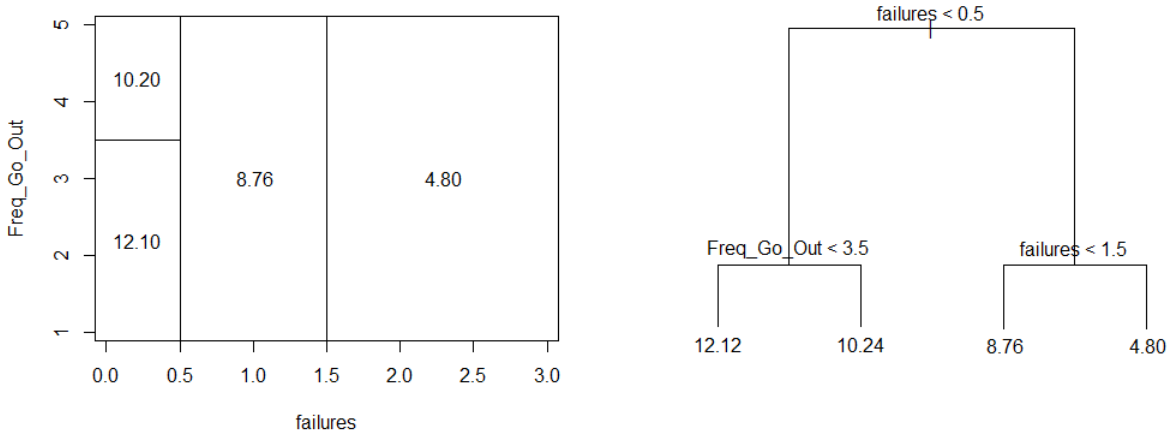
First of all, we need to distinguish between a regression tree and a classification tree: if the dependent variable is quantitative, then we are dealing with a regression tree, on the contrary, if the dependent variable is a qualitative one, we are considering a classification tree (therefore we are predicting the class in which the considered observation belongs).

The functioning of a CART is the following: it partitions the feature space into several non-overlapping regions, accordingly to a pre-determined algorithm. Each region established by this partition is a leave of the tree. In the case of a regression tree, inside any leave, each observation is predicted to have the same value. This value is determined as the average of the observations belonging to that leave in the training sample used to construct the space partition. On the contrary, when considering a classification tree, the prediction of an observation that falls inside a partition is given by the most commonly occurring class of the training sample in that leave.

In order to provide an example of the functioning of CARTs, we apply them to the Portuguese school dataset. Given that the dependent variable, the result of the exam, is numeric, we are dealing with a regression tree. For simplicity, we start taking into account only two characteristics of the students: the number of classes failed and the frequency of going out with friends. This second variables goes from 1 (rarely going out) to 5 (going out extremely frequently) and it is a self-reported measure. We now create a partition of the feature space to predict the final exam grade of a student based on these two characteristics. This partition is shown below in figure 1.3.

The left graph of this figure shows the regions in which the exams results are stratified. If a student has failed a class more than 1.5 times (i.e. two times or more) he is predicted to have a result at the exam of 4.8. This involve that in this partition (those who failed more than 1.5 times) of the training sample the average grade is 4.8. Moreover, all the students with this characteristic are predicted with the same exam result, independently of the frequency they go out with friends. On the contrary, when a student has failed less than 0.5 classes (i.e. he never failed one), his predicted outcome will depend also on the frequency of going out. If this variable is higher than 3.5, than the student is predicted to score a 10.2, if it is lower, he is predicted to score 12.1.

Figure 1.3: Different representations of the feature space partition (numeric dependent variable)



The graph on the left shows the feature space partition using the Portuguese school dataset, with the dependent variable being numeric. The graph on the right shows the same partition represented in tree form.

The right graph of figure 1.3 shows the same concept but using the tree representation. Using the tree representation to graphically report the partition of the feature space is probably more intuitive and we will therefore always use this method. Moreover, the regions representation is possible only with a number of variables equal or lower than three.

The graphs above presented the final outcome of a regression tree, however nothing has been said on the algorithm that partitions the feature space. The algorithm used will depend on which kind of tree we are considering. We start with a regression tree. In this case, the objective is to partition the feature space in high-dimensional rectangles that minimize the residual sum of squares (RSS). In order to make this process computationally feasible, the algorithm that is implemented in a regression tree follows a so-called recursive binary splitting, that James et al. 2013 define as a top-down and greedy approach. It is top-down because it starts to partition the feature space at the highest point of the tree, when all the observations are contained in a single leave. It is greedy, or myopic we could say, because at each step of the partitioning process, it selects the best split at that particular step, without considering the consequences of that split on the future splits. Using a parallelism with chess, this algorithm acts like a player that decides his move analysing only one move ahead and neglecting the consequences of his move in the development of the game.

The recursive binary splitting algorithm works by selecting a single feature and a single cut point of the selected feature such that, if the sample is divided in those observations that are above the cut point and below the cut point, the RSS is minimized. Formally, given a predictor X_j and a

cut point c, the recursive binary splitting algorithm creates the following two regions:

$$R_1(j,c) = \{X|X_j < c\} \; and \; R_2(j,c) = \{X|X_j \geq c\} \tag{1.2}$$

These two regions are designed in a way that the following quantity is minimized:

$$\sum_{i:x_i \in R_1(j,c)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,c)} (y_i - \hat{y}_{R_2})^2 \tag{1.3}$$

Where $\hat{y}_{R_1}$ and $\hat{y}_{R_2}$ represent the mean outcome in the two regions.

The above process is repeated until a certain stopping rule is verified. For example, the splitting process could continue until no leaves has more than n observations.

A relevant problem with this approach is that it can heavily overfit data, i.e. it could capture noise, instead of relevant characteristic that predicts the outcome. Because we are interested in out-of-sample predictive power, we want to reduce overfitting as much as possible. A first solution could be imposing a stopping rule that constraints the growth of the tree. However, given the nature of the algorithm, that tends to miss good splits that are in lower branches of the tree, this method often does not maximize predictive power. A better alternative is represented by tree pruning. The idea of this approach is to grow a large tree and then pruning those branches that do not improve significantly the RSS. As we will not use pruning in this thesis, we do not provide a more detailed explanation of this technique[6]. An alternative to pruning to reduce overfitting is considering various trees and averaging their predictions. These techniques are presented in the next section. Before moving to it, we first need to describe more in detail the functioning of a classification tree.

A classification tree is used when the dependent variable is qualitative. In this case, the prediction of an observation that falls inside a partition is given by the most commonly occurring class of the training sample in that leave.

We can implement a classification tree on the Portuguese school dataset if we consider the dependent variable as passed (grade higher or equal than 9.5) or not passed (grade lower than 9.5). In figure 1.4, we replicate the same analysis that we performed with regression tree using this new specification of the dependent variable (passed/ not passed).

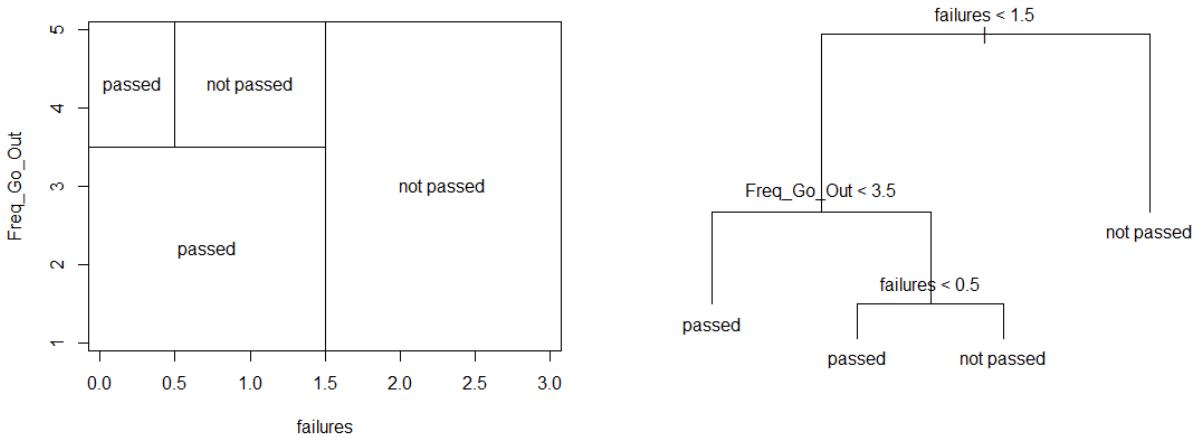
The two graphs show that those individuals who have failed previous classes more than 1.5 times are predicted not to pass the exams. This means that in the training sample, among those who failed more than 1.5 times, the most commonly occurring class is "not passed". Moreover,

[6]James et al. 2013 describes the process of pruning a tree in detail in chapter 8 of their book.

these students are predicted to fail regardless of their frequency of going out. On the contrary, when the variable failure is lower than 1.5 but higher than 0.5, the predicted class will depend on his frequency of going out. Finally, when a student has never failed a class previously, he is predicted to pass the exam. However, there is a distinction among these individuals that depends on the frequency of going out. Even if in both the partitions the predicted outcome is "passed", they differ in what is called node purity.

Node purity refers to the percentage of observations belonging to the most commonly occurring class. In our example it is likely that the group of those who go out often has a lower class purity than the group of those who do it less often. It means that the percentage of students who passed the exam in the first group is lower than the second group in the training sample. Class purity is important because it give us an idea of what is the probability that an individual will be correctly classified out-of-sample. Finally, it is important to notice that the feature space partition is different with respect to the regression tree.

Figure 1.4: Different representations of the feature space partition (binary dependent variable)



The graph on the left shows the feature space partition using the Portuguese school dataset, with the dependent variable being binary. The graph on the right shows the same partition represented in tree form.

A question that remains unanswered is how the splitting process is preformed, given that with qualitative variables it is not possible to use RSS as a measure of goodness of fit. The most straightforward way to adapt regression tree to the case of classification is to use the classification error rate. However, this criterion turns out to perform badly in split selections. For this reason the Gini index is preferred in classification tree[7].

[7] An alternative to the Gini index is cross-entropy, given by the formula $D = -\sum_{k=1}^{K} \hat{p}_{m,k}$.

The Gini index is given by the following formula:

$$\sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k}) \tag{1.4}$$

Where $\hat{p}_{m,k}$ is a measure of node purity, as it expresses the ratio between the number of observations in region m that belong to class k and the total number of observations in region m. Therefore, if $\hat{p}_{m,k} = 1$ it means that all the training observations in region m belong to class k. In the example of the Portuguese school data we have only two classes (passed and not passed) and therefore the Gini index for region m would be the following:

$$G = \hat{p}_{m,passed}(1 - \hat{p}_{m,passed}) + \hat{p}_{m,notpassed}(1 - \hat{p}_{m,notpassed}) \tag{1.5}$$

If in region m there is a high node purity, one of the two ratios will take a value close to one, while the other will be close to zero, determining a low value of the Gini index. The aim of the algorithm is exactly to find the split that minimize the Gini index.

1.2.3 Bagging

The CARTs described in the previous section suffered from overfitting, i.e. they could produce extremely good predictions in the training sample but had a weak predictive power in a hold-out sample.

We now present different approaches that allow to improve the prediction power using decision trees.

Bootstrap Aggregation[8], or Bagging, is a method based on the creation of a high number of trees. It produces an overall prediction resulting from averaging out the predictions obtained with the variour trees grown[9].

The idea underlying this method is that if we can grow various trees using different samples, and then average the results obtained, we will achieve a model that overfit data much less than a single tree. As we only have one training sample, we need to use bootstrapping to generate a number B of new samples that are used to grow B trees[10]. Once the trees are grown it is

[8]This machine learning algorithm was developed by Breiman 1996.

[9]The concept of bagging is broader than the one descripted. It refers to any method that uses bootstrapping to reduce the variance of a statistical method. However, in this thesis we will focus only on its application to decision tree that is by far the most used.

[10]While running bagging (and random forest) there is no concern about overfitting the data if B is very high. Indeed, what happens is the opposite. The higher B, the more reliable our predictions will be. Consequently, B should be as high as computationally feasible, even tens of thousands. Most of the applications that we can find in the economics literature use at least B=500.

sufficient to average their forecasts to obtain the bagging prediction[11].

Defined $\hat{f}^b(x)$ as the prediction of the bth tree, the formula of a bagging prediction is the following:

$$\hat{f}_{bag} = \frac{1}{B}\hat{f}^b(x) \tag{1.6}$$

The mechanism of bagging just described applied to quantitative variables, but it can be extended to qualitative ones. When implemented to a qualitative dependent variable, bagging algorithm follows the same procedure as above, growing a number B of trees after bootstrapping the training sample. However, it cannot produce an overall prediction as the average of the single predictions because these are categorical. The criterion adopted to produce a prediction is the one of majority vote. It means that the most commonly predicted class among the B predictions will be the bagging prediction.

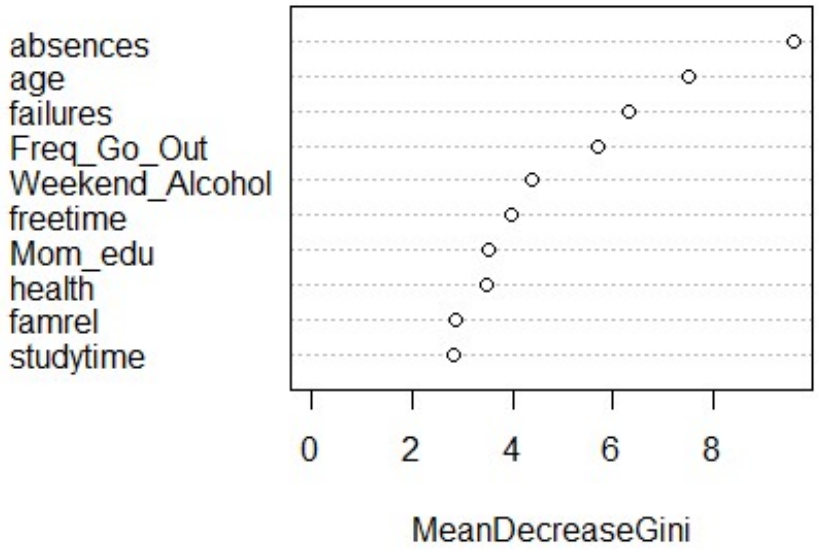
Bagging can drastically improve out-of-sample prediction accuracy if compared to a single tree. However, this improvement comes to a price, the loss of interpretability of the model[12]. When looking at a single tree we know what are the variables that influences the prediction and how they do it. In the example of the Portuguese school dataset, we knew that a student who already failed a class twice or more was predicted to fail the final exam. Using bagging (and other methods that combines various trees) it is not possible to make this kind of consideration, because we don't have information on how a variable influences the prediction. Nonetheless, it is still possible to have a measure of the importance of the variables used in the splitting process. Specifically, we can obtain a ranking of the mean amount, over the B trees, that the residual sum of squares is reduced because of splits over a certain variable. Alternatively, in the case of qualitative variable, we can compute the amount the Gini index is reduced by splits over a specific variable.

We can apply bagging to the Portuguese school dataset, with the qualitative dependent variable specification. As said above, the interpretability of the results is very low. However, it is possible to plot the variables importance as a function of the mean percentage points reduced in the Gini index. In the figure below only the ten most relevant variables are represented.

[11]These trees are grown deep, i.e. without pruning and with a stopping rule that allows the algorithm to create small leaves. The reason is that we are not worried about overfitting because it will be eliminated by the high number of trees averaged out.

[12]Another drawback of using bagging w.r.t. CARTs is that it can be computationally more expensive.

Figure 1.5: Variable importance plot using bagging



The graph shows the mean decrease in Gini index associated with each variable.

The meaning of figure 1.5 is that the feature *absences* (the number of absences during the scholastic year) is the one that on average reduces the most the Gini index. Specifically, the mean decrease in the Gini index, due to a split based on this variable is of almost ten percentage points. Other features that decrease on average the Gini index of at least five percentage points are age, $failures$ (the number of classed previously failed) and the frequency of going out with friends. These are the variables that have the highest influence in shaping the splits of the trees that compose the bagging model.

1.2.4 Random Forest

Random Forest, introduced by Breiman 2001, is a popular machine learning algorithm that can often improve the out-of-sample prediction quality w.r.t. bagging.

The way in which random forest can produce an improvement over bagging is by reducing the correlation between the trees that are used to produce the overall prediction. If we can reduce correlation between trees, than we will be able to reduce the overfitting of the data. Indeed, by producing a prediction that is the average of a large number of extremely similar trees, we reduce overfitting only modestly. On the contrary, if we can average out trees that are more different between them, we will achieve a more reliable forecast.

Now the question is how to decorralate trees. The intuition of random forest is that by using only a part of the full set of predictors it is possible to obtain trees that differ much more between

them. In particular, when random forest algorithm grows a tree, the number of variables that is considered to make any split is a randomly chosen subset m of all the variables. For example, in the Portuguese school dataset, there are 25 predictors, if we chose m=5, then when random forest is building a tree it will randomly select 5 variables out of 25 to perform every split. The 5 predictors randomly chosen change at each split.
The introduction of the parameter m, the number of predictors considered at each split, entails the need of using cross validation to determine the optimal level of this parameter[13]. It is worth remarking that if m is equal to the number of variables that is present in a dataset, then we are simply performing bagging.

1.2.5 Boosting

The third technique that can be implemented to improve the prediction quality of CARTs is Boosting. This machine learning algorithm is fundamentally different from the previous ones because it does not make use of bootstrapping, even if its aim is always to reduce overfitting.
The boosting algorithm works as follows: it initially grows a tree with d splits on the training data, then it computes the residuals between this initial prediction and the training data and it grows a new tree (with d splits) fitting the residuals instead of the outcome, finally it combines the initial tree with the new one (penalizing this last tree by a parameter λ). It repeats this process for a number B of trees.
To make clearer the functioning of boosting we report in the next page the way in which James et al. 2013 formalize this algorithm.
The rationale of boosting is to grow an initial tree that has low predictive power and subsequently slowly improving its performance fitting the residuals between the predictions of the model and the training data. One drawback of boosting is that, differently from random forest and bagging, it can overfit the data if the number of trees chosen is too high. For this reason, the number of trees has to be chosen by cross validation.
Generally speaking, cross validation plays a crucial role when implementing boosting because there are 4 parameters that must be chosen. The first one is the number B of trees. As we said, we want to find a value that is not too high because it would overfit the data, but we want it to be high enough to capture adequately the patterns that describe the outcome. The second parameter is the number d of splits that should be made in each tree. We should include among the values considered also d=1, i.e. each tree makes only one split. It might seem an extreme solution but it actually proves to perform well in different solution as showed by James et al. 2013. The

[13] James et al. 2013 suggests to use $m \approx \sqrt{p}$ where p is the total number of predictors in a dataset.

third parameter to be considered is the minimum number of observations in each node, that is particularly relevant when our training sample has not an elevated number of observations or if we chose a high value of d. Finally, the last parameter is λ, the shrinking parameters. Usually the values applied are smaller than 0.01, especially when B is high.

Boosting algorithm as formalized by James et al. 2013

---

1) Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all i in the training set (where r_i are the residuals, y_i the dependent variable and $\hat{f}(x)$ a feature space partition)

2) For $B = 1, 2, ..., B$ repeat:
i) Fit a tree $\hat{f}^b(x)$ with d splits to the training data
ii) Update $\hat{f}(x)$ by adding $\hat{f}^b(x)$ penalized by a parameter λ:
$\hat{f}(x) = \hat{f}(x) + \lambda \hat{f}^b(x)$
iii) Update the residuals r_i:
$r_i = r_i - \lambda \hat{f}^b(x)$

3) Output the final model:
$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x)$

As we have seen, the tuning process[14] is crucial when implementing a boosting model. However, the fact that we need to set various parameters could generate a computational problem. While implementing random forest, we need to choose only one parameter, the number of features to be considered at each split. Therefore, even if the number of variables is huge, let us say two thousand; the number of models that have to be implemented[15] is relatively low. For example, it would be 200 if we consider the sequence $m = \{1, 10, 20, \ldots, 1990, 2000\}$. This process of analyzing 200 models could last days if we are working with a big dataset, but it will be computationally feasible in most of the cases[16]. However, when implementing a boosting model, the different specifications can easily go to thousands. For example if we consider 5 values for B, the number of trees, 10 values for λ, the shrinking parameter, 10 values for d, the

[14]The process of choosing the parameters that maximize out-of-sample predictive power.

[15]We need to implement one model for each different setting of parameters that we are using, and then test with cross validation the prediction quality of this model.

[16]The cross validation process that we will illustrate in the second chapter indeed took a couple of days.

number of splits adopted by each tree, and finally 5 different values for the minimum number of observation in each node, the total number of models considered is 2500. Determining by cross validation the prediction quality of all these models will result difficult if we are working with big dataset because the time of computation could be huge. Indeed, in the second chapter of this thesis, a problem of this kind aroused when implementing boosting in our analysis. As a consequence, we had to reduce the number of parameters taken into account. However, this reduction could lead to a worse performing model as we could miss the best performing parameters.

1.3 Evaluating the performance of a model

In this section, we present the different techniques that are used in this thesis to evaluate the predictive power of a model. We focus on the performance evaluation in the case of a binary dependent variable, because this is the type of variable we are interested in our analysis.

The first thing to keep in mind, is that if we want to obtain a realistic estimate of a model predictive power, we need to use never-seen-before data. Different studies with simulated data prove that evaluating the prediction accuracy of a model on the same dataset that has been used to fit the model will overestimate its performance because of overfitting[17]. For this reason, in this section and in all the subsequent analysis that we will perform, the forecasts precision is always evaluated with a hold-out sample.

The most straightforward way to evaluate the quality of a binary prediction is by using classification accuracy, which quantifies the percentage of correctly predicted observations. This measure is of simple interpretation, but it can be deceiving. In general, we could think that a model with a higher classification accuracy is better than a model with a lower accuracy. However, it is not always the case. Let us consider a sample in which only 1% of the individuals suffer of a certain disease. Suppose we fitted two models, the first one can classify correctly 90% of the healthy individuals and 90% of the sick ones, resulting in a 90% classification accuracy. The second model simply assigns all the individuals to the healthy class, producing a 99% classification accuracy. If we had to judge only by using this measure of prediction quality, we would choose a model that has no predictive power against another that performs extremely well.

A more informative approach is to use a confusion matrix. In a confusion matrix, each column represents the predicted class of an observation, while each row represents the actual class of that

[17] James et al. 2013 provides different examples with simulated data of overestimation of the model performance using the same dataset for training and for testing. An alternative to a holdout sample is cross validation. It has the advantage that all the sample can be employed to train the model but it is a worse estimate of the real performance of the model as it tends to overestimates it (though much less than using the same sample for training and testing).

observation. With this matrix it is possible to understand not only the amount of misclassified observations, as it was possible with classification accuracy, but also to see what observations have been misclassified to which class. In this way, we have a clearer picture of the accuracy of a model.

In order to clarify this concept, we apply it to the example of the Portuguese school dataset. We provide below the confusion matrix resulting from OLS predictions on the train sample.

Table 1.1: Confusion matrix

Actual/Predicted	Not Passed (0)	Passed (1)
Not Passed (0)	22	38
Passed (1)	15	100

The confusion matrix of the OLS predictions on the Portuguese school dataset.

Table 1.1 is telling us that OLS model correctly classified 100 students that passed the exam and 22 students that did not pass the exam. However, it misclassified 15 students that were predicted not to pass the exam but that actually passed it, and 38 students that were predicted to pass it, but they did not.

The above example is useful to introduce the concepts of specificity and sensitivity. However, we need first to define the following notions:

- The true positives (TP) are those observations that are predicted to have passed the exam (or more generally, to belong in the class of interest[18]) and have actually passed it.
- The false positives (FP) are those observations that are predicted to have passed the exam but actually did not pass it.
- The true negatives (TN) are those observations that are correctly predicted not to have passed the exam.
- The false negatives (FN) are those observations that are predicted not to have passed the exam, but actually did.

Once these notions are known, we can move to the definition of specificity and sensitivity.

Sensitivity measures the ratio between the true positives and all the individuals with the characteristic of interest (TP+FN). In other words, this ratio tells us how a prediction of an individual

[18]In our analysis, the class of interest will be those individuals who smoke at the age of 16. In medicine, were these concepts are used frequently, the class of interest is usually represented by those patients with a certain medical condition or who responds in a certain way to a treatment.

belonging to the class of interest is reliable. The formula of sensitivity is the following:

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{1.7}$$

In our example, we are capturing the portion of individuals who passed the exam that were correctly predicted to do so by our model. Using the above confusion matrix, the sensitivity of this model would be 86.9%, because out of the 115 students who passed the exam, we correctly classified 100 of them.

Specificity measures the ratio between the true negatives and all the individuals without the characteristic of interest (TP+FN). This ratio tells us how the model is reliable in classifying an individual as not having the characteristic of interest. The formula of specificity is the following:

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive} \tag{1.8}$$

Applying again this concept to the Portuguese school dataset example, we are finding the ratio between those individuals who were correctly classified not passing the exam and all the individuals who did not pass the exam. Therefore, we obtain a specificity of 59.5% because out of the 37 students who did not pass the exam, 22 of them were correctly classified.

Specificity and sensitivity are the two measures that we will use in our study to evaluate the prediction quality of a model. Unfortunately, these two measures strongly depend on the threshold that we adopt to classify one individual as belonging to one group or another.

Classification with a binary dependent variable generally works as follow: an individual is assigned to the class in which he has the highest probability to belong, i.e. he is assigned to a class if he has a probability of 50% or higher to belong to that class. There are cases in which this threshold of 50% could not be appropriate. When one of the two classes represents a small portion of the sample, for example 1%, an individual with a predicted probability of belonging to the minority class of 40%, is 40 times higher more likely than the average to belong to the minority class, but if a threshold of 50% is used, then he is still classified in the majority class. Moreover, there are circumstances in which we are more interested in one class than another or when we are more worried about false negative than false positive. For all these reasons, we need to evaluate how a model performs with different thresholds, ideally with all of them. Receiver operating characteristics (ROC) curves are the instruments used to perform this kind of analysis and we see now how they work[19].

[19]For a comprehensive explanation of how ROC curves work see Fawcett 2006

A ROC curve plots sensitivity as a function of specificity, using different threshold levels. The algorithm that build a ROC curve works in this way: first, it ranks all the predicted probabilities[20] of belonging to the positive class from the lowest to the highest. Subsequently, it assigns all the predictions to the negative class. Sensitivity will therefore be zero and specificity will be one. Then, it classifies the observation with the highest probability as being in the positive class, and it computes again specificity and sensitivity. If this observation is correctly classified, specificity will remain one while sensitivity will modestly increase. On the contrary, if it is misclassified, sensitivity will remain at zero and specificity will decrease. The algorithm than selects the observation with the second highest probability and repeats the same process. It does the same with the third, fourth etc. observations until the one ranked with the lowest probability of being positive. When also this last observation is classified in the positive class, then sensitivity will be one and specificity zero. Finally, the levels of sensitivity and specificity obtained at each step of the algorithm are plotted.

In figure 1.6, we plotted two ROC curves obtained from the predictions on the test sample of the Portuguese school dataset. The two models implemented are OLS (in black) and boosting (in red).

How to interpret this graph? The 45 degrees gray line represents randomness. The closer a ROC curve is to this line, the worse its predictive power, the further, the better predictive power. However, it is still not clear from the graph if boosting is performing better than OLS.

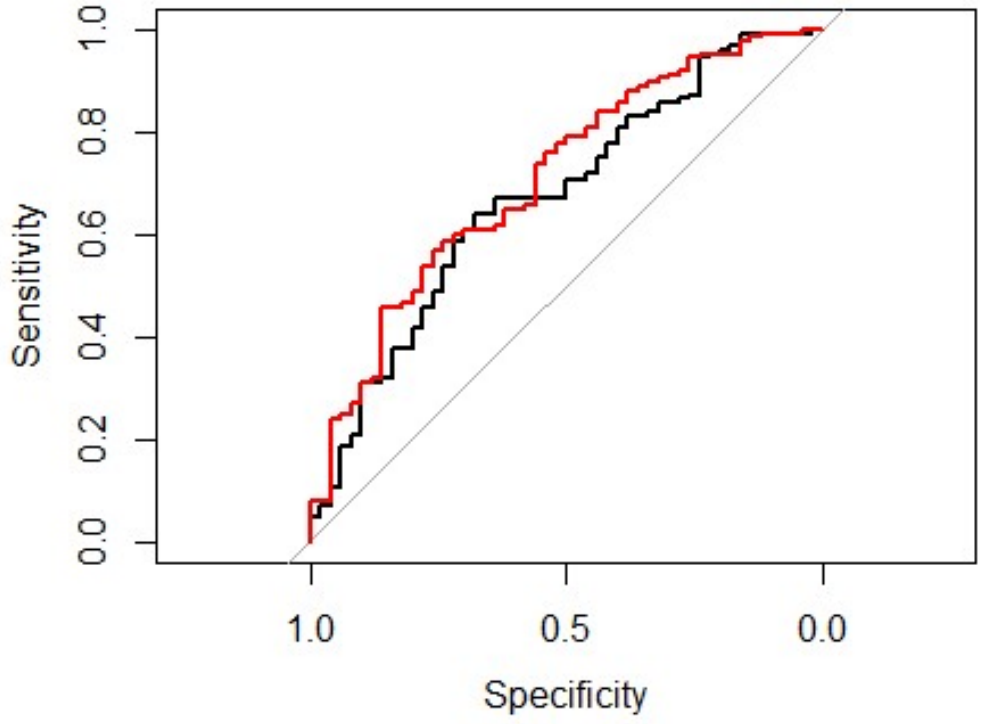
The criterion used to asses a ROC curve is the Area Under the Curve (AUC). It can vary from 0.5, when the model is not doing any better than random, to 1, when the model is perfectly ranking observations. The AUC of a ROC curve has an important statistical property: it is equal to the probability that a randomly chosen instance belonging to the positive class is ranked higher than a randomly chosen instance belonging to the negative class.

The AUC obtained with boosting is 0.70, while the one with OLS is 0.67. Therefore, the second is performing slightly worse than the first on the test sample used. In order to achieve reliable results, we need to construct confidence intervals for these values. We can create them bootstrapping the testing sample. In this way it is possible to compare different models with a method that provides a measure of prediction quality and confidence intervals for it.

We report in figure 1.6 the AUC obtained with the different methods that we used on the Portuguese dataset.

[20]We do not need these probabilities to be bounded in a zero one interval because the algorithm used to build ROC curves is based only on the rank of the predictions instead of their absolute values. For this reason, if we use OLS to predict probabilities we do not need to make any arrangement to the predicted values, even if there are negatives and more-than-one predicted probabilities.

Figure 1.6: ROC curves



ROC curves of boosting (in red) and OLS models (in black) applied to the Portuguese school dataset.

Table 1.2: Comparison of AUC using different models

Model	**AUC**
OLS	67.14%
	[58.58%, 74.90%]
Random Forest	66.46%
	[58.10%, 75.04%]
Boosting	68.91%
	[60.65%, 77.43%]
LASSO	69.14%
	[60.53%, 78.46%]
Ensemble	69.09%
	[60.14%, 77.86%]

The table shows the Area Under the Curve obtained implementing the different models considered in our analysis. In the parenthesis the bootstrapped confidence intervals are reported.

The first think that we want to verify is if a model is preforming statistically significantly better than random. The confidence intervals reported in table 1.2 exclude that the AUC of the models considered is equal to 0.5, then they are performing better than random. If we compare the performance of different models, we see that it is not possible to say that any of them is performing better than another as their confidence intervals always overlap.

Having presented ROC curves, the review of the machine learning techniques and methods to evaluate predictions' quality adopted in this thesis is completed. In the next chapter these methods are put into practice to predict adolescents' smoking habits.

Chapter 2

Application 1: Predicting smoking habits

2.1 Background and literature review

"The tobacco epidemic is one of the biggest public health threats the world has ever faced, killing more than 7 million people a year. More than 6 million of those deaths are the result of direct tobacco use while around 890,000 are the result of non-smokers being exposed to second-hand smoke"

2017 WHO report on the global tobacco epidemic

These dramatic figures underline that the reduction of cigarettes consumption is a compelling challenge for our society. A crucial aspect of it is to understand the patterns that drive tobacco use to help policy-makers designing targeted and effective policies against smoke. The contribution of this work is to propose a new instrument to help preventing smoking onset among adolescents[1]. We begin our study by summing up the main findings of the literature on the determinants of smoking inception among the youngest. Wellman et al. 2016 reviewed 53 articles that analysed the drivers of smoking onset and found various predictors that are statistically significant in a consistent number of studies. An increased probability of starting to smoke is explained by age, lower socio economics status, poor scholastic performance, sensation seeking, rebelliousness, intention to smoke in the future, friends and family members' smoking habits, receptivity to tobacco advertisement, and exposure to films. On the contrary, high self-esteem and parental monitoring explain in a statistically significant way a decrease in the probability of starting to smoke. Sex and ethnicity, even if statistically significant in different studies, are associated with contrasting evidences.

[1]The focus on adolescence is also justified by a strong evidence that daily smoking during this period of life is highly linked with smoking during adulthood (Saddleson et al. 2016) and with a lower probability of quitting (Chen, Millar, et al. 1998).

Parent practices (e.g. smoking-related communication and monitoring) play a central role in determining the smoking habits of a teenager. Thomas et al. 2015 provide a systematic review of the literature on this topic. The authors find out that a strict smoking-ban is an effective strategy to contrast the inception of smoking among adolescents. Moreover, the authors observed that a consistent number of studies provided evidence that reducing cigarettes availability at home has a preventive effect. Concerning the roles of communication, parenting reactions and parental norms, they found contrasting results. Some studies prove a significant effect of these parent practices to reduce the risk of smoking inception, but other papers do not find a significant impact. Finally, non-smoking agreements proved to be ineffective as no study could find a significant effect of this practice after controlling for the relevant factors.

The above-mentioned articles show that the determinants of smoking among the youngest are already well known. For this reason, we propose a new approach for analysing smoking onset among adolescents that is fundamentally different from the existing literature. Instead of looking for significant coefficients, that demonstrate a causal relationship between a certain variable and an increase in the probability of smoking, we will concentrate our effort on creating a model that maximise prediction accuracy of the future smoking status of an adolescent. In other words, the aim of the following analysis is to identify a group of teenagers who have a high probability of starting to smoke in the subsequent years. Being able to identify such a group, makes it possible to correctly target a policy intervention to prevent smoking inception.

This approach is inspired by Kleinberg et al. 2015 that suggest how predictions are sometimes more adequate than inference to tackle a policy intervention problem. For example, Chandler, Levitt, and List 2011 create a model to single out a group of students with a high probability of being involved in shootings in the future. Once these students are identified, a mentoring approach can be efficiently targeted, and the risk of shootings reduced. Similarly, in this thesis, we will create a model that predicts the smoking status of a 16 years old adolescent based on his characteristics when he is 10 years old. The dataset used to carry out this analysis is the British Cohort Study of 1970 that is described in the next section.

Clearly, in order to apply this approach to a real policy problem, our model would need to be tuned with the available data, as the input of the model is specific of the contest in which it is applied. However, the results obtained in terms of predictive power are promising and suggest that this approach could be successfully implemented in a real policy application.

From a methodological point of view, the techniques implemented to develop this analysis are the ones presented in the previous chapter, i.e. LASSO, bagging, random forest and boosting. The application of these techniques is extremely limited in the field of smoking prevention. Frank, Habach, and Seetan 2018 make use of machine learning to classify the current smoking

status of an individual, based on blood tests and vital readings data. Dumortier et al. 2016 try to predict urges to smoke during a smoking quit attempt, implementing discriminant analysis and decision tree learning models[2]. Their results are interesting from different point of views, but they do not contribute to our purpose of predicting and preventing future smoking.

2.2 Data

Our analysis is based on the British Cohort Study dataset of 1970 (BCS70). This longitudinal study collects a wide range of features from a sample of over 17,000 babies born in England, Wales, Scotland and Northern Ireland during a single week of 1970. Since the first survey, that was carried out at birth in 1970, there have been other nine sweeps that collected respondents' data in the years 1975, 1980, 1986, 1996, 2000, 2004, 2008, 2012 and 2016. The 2016 wave results will be available only from 2019, while the next sweep is scheduled for 2020.

The waves of our interest are the ones that took place in 1980 and in 1986, when the cohort members were respectively 10 and 16 years old. However, the subsequent waves also provide precious information for our analysis, because we can observe how the smoking habits of an individual evolved during his life.

The target of our predictions, the dependent variable, is the self-declared number of cigarettes smoked in a week at age 16[3], taken from the 1986 sweep. Unfortunately, not all the cohort members answered the document C of the survey that includes this information on smoking habits. Consequently, out of the 11,622 respondents who participated at the fourth wave of the BCS70, we dispose of the self-declared number of cigarettes smoked for only 5,450 individuals. The distribution of this variable is reported in the table below.

Table 2.1: Original distribution of the dependent variable

Smoking habits	**Freq.**	**Percent**	**Cum.**
None	4,269	78.33 %	78.33 %
Less than one	153	2.81%	81.14%
Between one and four	149	2.73%	83.87%
Five or more	879	16.13%	100.00%
TOTAL	5,450	100.00%	

[2]Moreover, Suchting et al. 2017 implement an elastic net penalized cox proportional hazards regression with the aim of predicting smoking relapse.

[3]The reliability of this self-declared measure is discussed later in the chapter.

For the purpose of our analysis, this variable is coded as a dummy, taking value 1 if an individual is a smoker and value 0 otherwise. Those who revealed to smoke less than one cigarette per week and between one and four cigarettes per week are dropped from the sample. The choice of this specification is driven by the evidence that a strong increase in the probability of persistent smoking during adult life relates to regular smoking in adolescence instead of occasional smoking[4]. The new distribution of the dependent variable is reported in figure 2.2.

Concerning the predictors used in the model, they all refer to the children at age 10. This condition is crucial, because we want our model to be able to predict future smoking status without any information about the future conditions of the individual, as it would be in a real policy implementation.

Table 2.2: Distribution of the dependent variable coded as a dummy

Smoking habits	**Freq.**	**Percent**	**Cum.**
Non-smoker (0)	4,269	82.93 %	82.93 %
Smoker (1)	879	17.07%	100.00%
TOTAL	5,148	100.00%	

These predictors cover a broad range of characteristics of the individuals, including most of the features considered by the literature as relevant in determining the smoking habits of an adolescent. We inserted in the model 148 variables that relates to the following areas [5]: basic demographic information, family background, parents, relatives and friends' smoking habits, personality traits, leisure activities and cognitive skills.

Unfortunately, the BCS70 present a relevant problem of missing values. Out of the 5,148 observations of which we know the smoking status at age 16, only 2404 of them do not present missing values in the variables of interest [6]. As a consequence, only this smaller subsample can be examined in our study.

[4]For example as shown in Saddleson et al. 2016.

[5]The number of variables includes various dummies needed to implement qualitative characteristic. A different specification of the model is proposed in appendix B.

[6]This subsample is obtained in a non-random manner, because it represents those individuals who answered all the questions included in our model. The fact of answering all the questions could be linked to some individual characteristics (e.g. income or school performance). Therefore, any inference analysis would be biased. Being our aim to create a predictive model, this sample selection still influences its external validity. The patterns that describe the link between the predictors and the predicted outcome could be different in this sample and in the population considered by the British Cohort Study. However, we are mainly interested in evaluating if it is possible to predict future smoking status and evaluate the performance of the different models, instead of creating a model that is applicable to the population considered by this survey. Therefore, we should not be too concerned about this bias. Moreover, any implementation of this model to a real policy problem would require to re-tune the model on the sample of interest, as the characteristics available would surely differ from the ones considered in our analysis.

This relatively high level of unanswered questions strongly affects the potential of our analysis. An interesting characteristic of the BCS70 is that it includes hundreds of features in each wave. Machine learning could perform a variable selection identifying the most relevant for the predicted outcome.

Unfortunately, due to missing values, it is not possible to perform this variable selection, because the number of observations for which we have a full set of information is almost zero.

Therefore, we needed to select manually the variables to be included in the model [7], but we could have missed important ones to maximise prediction accuracy.

We now provide some descriptive statistics of the subsample of 2404 individuals that will be used for our analysis. First of all, we report in the table below the distribution of smoking at age 16 in this sample.

Table 2.3: Distribution of the dependent variable in the sample object of our analysis

Smoking habits	**Freq.**	**Percent**	**Cum.**
Non-smoker (0)	2022	84.11 %	84.11 %
Smoker (1)	382	15.89%	100.00%
TOTAL	2404	100.00%	

In the next table we compare smokers and non-smokers at age 16 with respect to some key characteristics at age 10. We also provide the level of significance of the differences in the features examined between the two samples.

The two groups differ in almost all of the characteristics considered, most of them consistently with what we would expect from the literature. The features included in table 2.4 refer to the individual at age 10, while the smoking status is measured at age 16.

The results of this comparison show that the smokers group is significantly poorer than the non-smokers one. The family income of the latter is lower than 50 thousand dollars in 5.76% of the cases, opposed to a 2.52% for the first group. Smokers parents' cigarettes consumption (on average 7.39 daily cigarettes for the mother and 9.98 for the father) is significantly higher than non-smokers parents' consumption (on average 4.38 daily cigarettes for the mother and 5.81for the father). Moreover, females represent a significantly higher percentage (almost 64%) of the smokers sample than the non-smokers one (56%).

[7] In this selection we followed the existing literature to include in the model most of the characteristics considered relevant in the onset of smoking among adolescent.

Table 2.4: Comparison between smokers and non-smokers

Variable	**Entire sample mean**	**Non-smokers mean (a)**	**Smokers mean (b)**	**Level of significance of a-b**
Percentage female	57.32%	56.08%	63.87%	***
Mother's cigarettes daily	4.86	4.38	7.39	***
Father's cigarettes daily	6.47	5.81	9.98	***
Frequency friends smoke	0.30	0.32	0.22	***
Math test result	48.16	48.53	46.18	***
Smoke damage health	0.099	0.098	0.102	
Fights with other children	15.25	14.60	18.68	***
Impulsive	28.84	27.89	33.81	***
Often disobedient	22.46	21.42	27.92	***
Frequency going to cinema	1.61	1.59	1.70	**
Frequency Sport	2.42	2.41	2.50	**
Percentage income <50k	3.03%	2.52%	5.76%	***

Comparison between smokers and non-smokers characteristics. The last column reports p-values of the differences between the two groups: *p<0.10; **p<0.05; ***p<0.01. The frequency of friends who smoke is a self-reported measure that takes value zero when none of the child's friend smoke at age 10, value one when some of them smoke, and value two when most of them smoke. The math test results vary from 0 to 72. The variable "smoke damage health" take value zero when the child agrees on the fact that smoking damage health, value 1 if he partially believe it and value 2 if he does not believe it. The variables "fights with other children", "often disobedient" and "impulsive" are measured in a scale from zero (do not apply) to one hundred (completely apply). These measures are reported by the child's mother. Also the frequency of going to the cinema and of playing sports are reported by the child's mother and take value one if the child never does the activity, value two if he does it sometimes and three if he does it often.

Concerning academic performance, the adolescents who make use of tobacco at age 16 performed significantly worse in a math test taken at age 10.

Looking at the children behavior, the smokers used to be more disobedient, impulsive and fight more often with other children than the non-smokers. Finally, the adolescents who make use of tobacco at age 16 used to go more often to the cinema and to play more frequently sports than those who do not smoke. All these results are mostly consistent with the existing literature, even if clearly no causal relation can be inferred from this comparison.

The only result that is different from our expectations is the one on the belief about the impact of smoking on health. No difference between the two groups has been observed: in both groups the vast majority of children is aware of the damages that smoking has on health.

To conclude this focus on the data, we study how smoking habits of the 2404 adolescents included in this analysis evolved over time. So far, we concentrated on the information available at ages 10 and 16. However, the British Cohort Study follows the individuals up to the age of 42, giving us the opportunity to see how their smoking habits has changed over time. The two tables below report the percentage of individuals who consume a certain quantity of cigarettes, sorted by their smoking status at age 16. The first table reports data at age 34, while the second at age 42. The slightly difference of specification in cigarettes consumed (at age 34, a category "occasional smoker" is included, while at age 42 it is not), derives from the way in which the data were collected in the two waves.

Consistently with our expectations based on the literature, we observe that those adolescents who used to smoke daily at age 16 are more likely to be frequent smokers at later stages in their lives. At age 34, among the group of non-smokers during adolescence, only around 10% of the individuals is a regular smoker. On the contrary in the group of smokers, this percentage increases to more than 50 %. Even if slightly less significant, this gap between smokers and non-smokers is maintained at age 42.

These findings are extremely relevant for our analysis. They confirm that smoking habits are persistent during the lifetime of an individual. In particular, they show that if an individual is not smoking daily during adolescence, he is very unlikely to start to smoke later on in his life. As a consequence, focusing on prevention during adolescence seems an effective strategy to reduce tobacco consumption in the population of any age. Indeed, if we can prevent a child to start to smoke during his adolescence, we are confident that we are drastically reducing his chances to smoke during his whole adult life.

Table 2.5: Smoking habits at ages 34 and 42

Age 34					
	Non-smoker	**Occasional smoker**	**Up to 10**	**Up to 20**	**More than 20**
Non-smoker (0)	84.57%	5.05%	4.79 %	4.92 %	0.64%
Smoker (1)	38.35%	8.96%	21.51%	29.03 %	2.15 %
Age 42					
	Non-smoker	**Occasional smoker**	**Up to 10**	**Up to 20**	**More than 20**
Non-smoker (0)	92.06%	/	3.50%	3.77 %	0.66%
Smoker (1)	61.02%	/	14.71%	20.96%	3.31 %

Smoking status at age 16 is reported in rows. The columns represent smoking habits at ages 34 and 42. The class "occasional smoker" is not included in the data collection at age 42.

2.3 Econometric strategy

The primary aim of this work is to create a model that correctly classify smoking status of adolescents at age 16 based on their characteristics at age 10. A strictly connected issue is to evaluate how accurate the outcome of our model is. It would be somehow useless to create a model without knowing if its performance is positive or not [8]. In order to carry out this kind of analysis we need to sort the observations available in two subsamples. The first (train sample) is used to fit the model. The second (test sample) is used to assess the quality of the model's predictions.

As we have seen in the first chapter, this process of splitting the data is crucial if we want to obtain a realistic estimate of the model's performance. If we asses a model on the same data on which we fitted it, then we could strongly overestimate its predictive power because of overfitting. Given that we are interested in the performance on never-seen-before data, that best reflects what would be the predictions' accuracy in a real policy implementation, we will evaluate performance exclusively out of sample.

Following Kleinberg et al. 2015, we assign to the train sample about two third randomly chosen observations and to the test sample the remaining one third. The choice of having a bigger

[8]Being this study the first that tries to carry out such an analysis, it is not easy to evaluate the performance of a model because there are no benchmarks to confront with. However, we want at least our model to perform significantly better than random (and we will test if it can do it).

training sample seems logic given that the whole sample considered is not extremely large and we want to prioritize the model fitting process instead of the comparison between different machine learning approaches.

The outcome of the model is the smoking status at age 16, coded as a dummy variable as shown in the previous section. The predictors used to create the models are the 148 features at age 10 described in the previous section.

The models implemented are the ones seen in the first chapter, i.e. LASSO [9], boosting, random forest, bagging and ensemble [10]. Moreover, we will also run a simple OLS regression to compare the performance of a traditional econometric tool with machine learning models.

Three of the models implemented (LASSO, random forest, boosting) need a process of parameters tuning in order to maximise their predictive power. We will see now the tuning strategy followed for each of these methods [11].

Concerning LASSO, the parameter to be tuned is λ. We created 100 different models making the value of λ moving from 0.1 to 0.001 with a decrease of 0.001 for each new model implemented. Subsequently, we evaluated the performance of each model with a 5-fold cross validation. The criterion adopted to assess the predictive strength of a model is classification accuracy, i.e. the percentage of correctly predicted observations [12]. The model that provided the highest classification accuracy has $\lambda = 0.025$, and therefore it is model that we implemented on the whole training sample.

Regarding random forest, the parameter to be tuned is m, the number of variables that are randomly chosen at each split of the tree growing process. We created 148 models, considering therefore all the possible values of m, from 1 to 148 [13]. We used again 5-fold cross validation and evaluated the models with classification accuracy. The value that produced the best model is m=55.

Finally, we had to tune the boosting model. This process is more challenging because the number of parameters to be chosen is 4. In this case a computational problem arises. If we try ten different values for each parameter, we would obtain 10^4 models that need to be evaluated by cross validation.

[9] Before running the LASSO regression, we created interaction terms between all the variables included in the model. With this specification the total number of variables considered is 10880.

[10] The ensemble model is simply created as an arithmetic average of the predicted probabilities of smoking of 3 machine learning models implemented (random forest, LASSO, boosting).

[11] The tuning process is carried out using the R package caret (Kuhn et al. 2008), that enables the user to perform cross validation with a wide range of machine learning techniques using the same command.

[12] Classification accuracy is not the most appropriate way to evaluate the quality of a model as we have seen in the first chapter. However, using AUC would be much more challenging both from a computational and implementation point of view.

[13] Notice that when m=148, the total number of features in the dataset, we are actually performing bagging. For this reason, we will not include bagging as a model because we will consider it as a special case of random forest.

Implementing such a high number of models is not computationally feasible. As a consequence, we needed to make a careful selection of the values to be included in our analysis. In order to make this selection we followed the suggestions of James et al. (2013). The table below reports all the parameters implemented (overall, 144 models).

Table 2.6: Boosting cross validation parameters

Parameter considered	**Values implemented**
Interaction depth	1,5,10,50
Number of trees	500, 1000, 5000
Shrinkage parameter	0.1, 0.01, 0.001, 0.0001
Minimum node size	3,15,40

The table shows the parameters implemented in the cross validation process of the boosting model.

The model that produced the best 5-fold cross validated classification accuracy is the one with 5000 trees, 10 splits per tree, 3 observations as the minimum node size and a shrinking parameter $\lambda = 0.001$.

After tuning the parameters of the models that needed this procedure, we could finally apply them to the whole training dataset and compare their performance.

2.4 Results

In this section we present the results of our analysis, concentrating our attention on the quality of predictions produced by each model previously described.

First of all, we focus on the respondents that are identified by the different models as the 10% most at risk of starting to smoke. We want to see what percentage of them actually start regular smoking at age 16.

This approach is particularly informative, as the aim of this research is exactly to help policy makers targeting with an intervention a selected group of individuals. Therefore, implementing this kind of analysis, we can quantify the percentage of potential smokers that would be targeted by a selective policy intervention.

We should keep in mind that in the sample considered, only 16% of the respondents will start to smoke at age 16. Consequently, targeting the right adolescents could drastically increase the efficiency of an intervention and improve the resources allocation to those people who would benefit the most from it.

We present in the table below the percentage of actual smokers at age 16 among the individuals predicted to be the 10% most at risk of starting to smoke[14]. In order to compare the performance of the different models we created bootstrapped confidence intervals by resampling 2000 times the initial sample.

The ensemble method outperformed in a statistically significant way all the other techniques, with the only exception of boosting. Ensemble can identify a group where the percentage of smokers is around 49%, that means more than 3 times the portion of smokers in the entire sample. This improvement in selecting teenagers who will start to smoke at age 16 could help policy makers to screen for those individuals who need the most a smoking prevention program.

Table 2.7: 10% most at risk performance comparison

Method used	**Percentage of actual smokers**
OLS	41.77%
	[40.03%, 43.51%]
Random Forest	42.85%
	[40.72%, 44.99%]
Boosting	48.10%
	[45.83%, 50.37%]
LASSO	43.04%
	[41.44%, 44.63%]
Ensemble	49.37%
	[45.55%, 53.23 %]

The table reports the percentage of actual smokers in the group of individuals identified by the different models as the 10% most at risk of smoking (in the hold-out sample). In parenthesis the bootstrapped confidence intervals are reported.

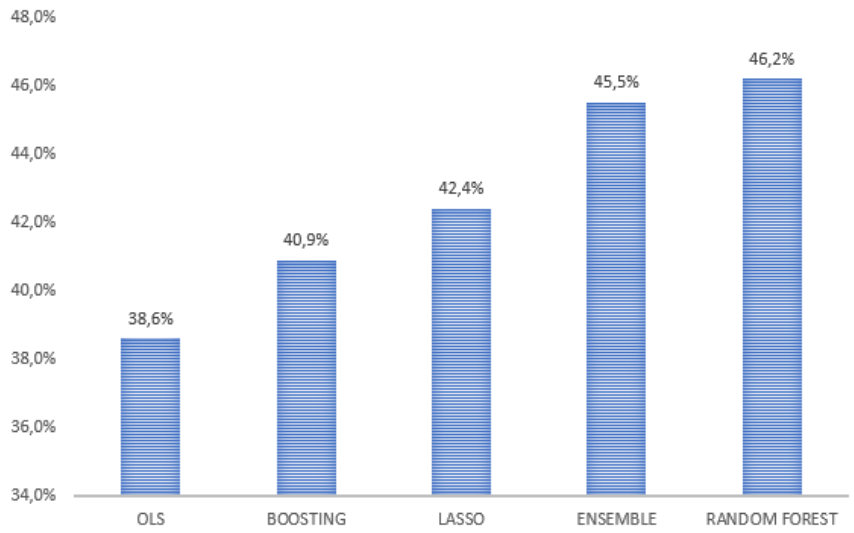
So far, we investigated the capacity of the models to correctly identifying a group with a high percentage of smokers at age 16. Instead, in the following paragraphs we study more generally the predictive power of the models considered using the approaches presented in the first chapter. For starters, we compute the level of sensitivity of the models' predictions on the test sample, i.e. the ratio between the true positive (i.e. the smokers correctly classified) and the sum of true positive and false negative (i.e. the smokers correctly classified plus the smokers wrongly classified). The threshold used to classify observations was chosen to keep specificity constant

[14]We can use different threshold than 10%. The performance of the models is assessed with different threshold in appendix A.

at an 85% level [15].

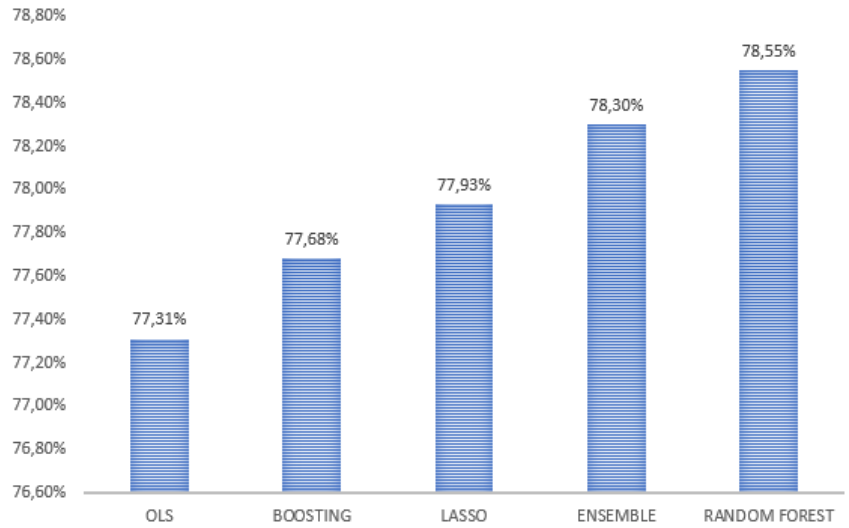
Figure 2.1 plots the sensitivity obtained on the test sample implementing the different models. Machine learning methods performed better than OLS. The latter produced a sensitivity of 38.6%, while the best performing ML model, random forest, reached a 46.2% level of sensitivity.

Figure 2.1: Sensitivity in the hold-out sample



The graph reports the level of sensitivity obtained in the hold-out sample with the different models considered.

Figure 2.2: Classification accuracy in the hold-out sample



The graph reports the classification accuracy obtained in the hold-out sample with the different models considered.

[15] We decided to focus on sensitivity and keeping specificity constant because we believe that the first measure is the most interesting for our purpose. Indeed, it gives us information on how good our model is in correctly classify a child who will start to smoke by the age 16. A model that can correctly identify these individuals is exactly what a policy maker would need to efficiently target its intervention.

Figure 2.2 reports the classification accuracy obtained with the different models. We need to consider that we are keeping specificity constant and that the positive class represents only 16% of the sample. Therefore, the small differences shown by the table are mainly explained by these two factors.

The classification accuracy obtained with OLS is 77.31%, while with random forest we achieve a 78.55%. Therefore, in terms of classification accuracy, if we want to keep specificity constant at 85%, the differences between the models are modest. Moreover, someone could argue that the classification accuracy is lower than what we would obtain by classifying all the individuals as non-smokers. Indeed, we would obtain a classification accuracy of 84%, being the smokers only 16% of the sample. However, as we have already seen in the first chapter, when one class (the non-smoker) is much more numerous than the other (the smoker), the level of classification accuracy obtained by assigning all the observations to the most numerous class can be higher than the one of an extremely informative model.

For example, a model that classifies smokers and non-smokers with a specificity and a sensitivity of 83% would have a classification accuracy of 83%, a lower value than the model that assign all the observation to the most numerous class. Nonetheless, its performance would be very positive.

Analysing sensitivity, specificity and classification accuracy in a hold-out sample, as we have done, is an approach commonly used in the literature (for example in Frank, Habach, and Seetan 2018 or in Chung and Li 2017). However, we believe that this approach presents an important weakness. These results are produced without confidence intervals. Therefore, we cannot say whether a model is performing significantly better than another.

To solve this problem, we created confidence intervals by bootstrapping the hold-out sample 2000 times. We focus on sensitivity measures because, as we have seen, the differences in classification accuracies are not particularly informative when keeping specificity constant.

In table 2.8, we report the bootstrapped confidence intervals and the original statistics obtained in the hold-out sample.

Looking at confidence intervals we observe that no method can perform significantly better than another in terms of sensitivity, given specificity at 85%. Therefore, even if in the test sample the differences in sensitivity seemed relevant, we cannot exclude the possibility that these differences derived from the specific sample drawn.

Moreover, a limit of the results presented above is that they are conditioned on a certain level of specificity, that is subjectively chosen. As we want to have a broader picture of the performance of the model at different specificity level, we will use ROC curves and the Area Under the Curve (AUC) to assess the predictive power of a model. Figure 2.3 shows the ROC curves obtained

with the predicted probabilities produced by each model and the actual outcome of the test sample.

The ROC curve of OLS model has a lower AUC than all the other curves, confirming that on the hold-out sample machine learning is performing better. In order to have statistically significant results of the models' performances, we provide in table 2.9 the AUCs obtained and the respective confidence intervals computed bootstrapping 2000 times the test sample.

The area under the curve of the OLS ROC curve is at 66.73%, while the best performing method, ensemble, produces an AUC of 73.69%[16]. All these methods perform significantly better than random, as their bootstrapped confidence intervals do not include the value 50%. However, we cannot say that any method is performing in a statistically significant way better than another, because confidence intervals overlaps.

The small test sample size could influence the confidence intervals width and prevent us from taking statistically significant conclusion. We tried a different sample slitting strategy, in which half of the sample was used to fit the model and the other half for testing[17] but we obtained again non-significant differences.

Table 2.8: Sensitivity (with confidence intervals) obtained in the hold-out sample

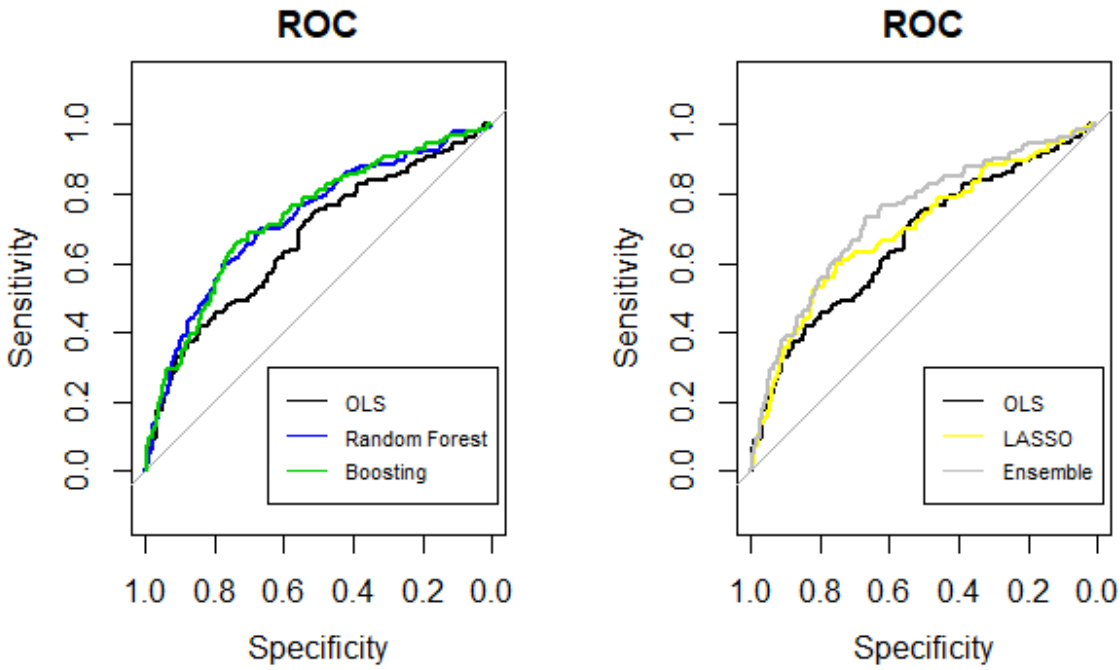
Method used	**Sensitivity level**
OLS	38.62%
	[30.30%, 48.48%]
Random Forest	46.24%
	[37.12%, 55.91%]
Boosting	40.92%
	[32.58%, 52.27%]
LASSO	42.42%
	[33.33%, 53.05%]
Ensemble	45.45%
	[36.36%, 55.30%]

The table shows the level of sensitivity obtained with the different models implemented. In parenthesis the bootstrapped confidence intervals are reported.

[16]This means that the probability that a randomly selected smoker at age 16 will have a higher predicted probability of being a smoker than a randomly selected non-smoker at age 16 is 73.69%.

[17]Therefore increasing the test sample size

Figure 2.3: ROC curves obtained in the hold-out sample



The graph on the left compares the performance of boosting and random forest with OLS, while the graph on the right compares the performance of LASSO and ensemble with OLS.

Table 2.9: Area Under the Curve (with confidence intervals) obtained in the hold-out sample

Method used	**Area Under the Curve**
OLS	66.73%
	[61.14%, 71.7%]
Random Forest	72.62%
	[67.46%, 77.57%]
Boosting	73.02%
	[67.75%, 77.41%]
LASSO	69.61%
	[63.95%, 74.39%]
Ensemble	73.69%
	[69.03%, 78.58%]

The table shows the AUC obtained with the different models implemented. In parenthesis the bootstrapped confidence intervals are reported.

2.5 Limitations and further research

The predictive power achieved with these models is promising, but probably not strong enough to qualify them as reliable instruments for identifying future smokers. However, our work can be extended and improved in different ways that we present in this section. By adopting some of the suggestions that we are proposing, we expect the models' predictive accuracy to increase, making them more powerful for targeting policy intervention. Nonetheless, this work represents an important starting point for further research because the methodology implemented can be replicated in future studies and the results obtained can be used as a benchmark to evaluate future results achieved.

Data A first relevant improvement could derive from the use of different data. The British Cohort Study has the advantage that it provides longitudinal data on smoking habits, together with hundreds of variables that describe the individuals' characteristics. Unfortunately, the high number of missing values in the sample, forced us to select a subset of the available variables to be considered for training the models. This selection, even if driven by the relevant literature, is likely to decrease the predictive power of machine learning models because we probably omitted features that are linked with smoking habits in adolescence. If we could work on a dataset with limited missing values, we would be able to let the machine learning algorithm select the features that are more relevant for predicting the outcome, without any bias introduced by our previous conception of the phenomenon[18].

Another important aspect to be considered is that the sweeps of the BCS70 taken into account in our analysis refer to the 80s. The patterns that determine smoking onset among adolescent are likely to be changed in more than 30 years. For example, it would be meaningful to include information on social network activity to study their predictive power on smoking habits[19].

The last relevant aspect related to the dataset used is the nature of the dependent variable. In the BCS70, smoking habits at age 16 are self-reported by the individual. Therefore, we could be worried that a portion of the smokers are not stating the truth on their smoking status. Moreover, the characteristics of those who say the truth and those who lie could differ, creating a bias in our prediction. In other words, we could create a model that identifies only the individuals who smoke and are willing to reveal the truth instead of the entire group of smokers.

[18] A relevant issue in the considered sweep of the British Cohort Study is that many individuals did not answered at all one or more of the questionnaires proposed. Consequently, the number of missing values is very high (20%-30% of the variables taken into account) for these respondents and imputations of missing data is not advisable.

[19] An interesting approach would be studying the activity of an individual on social network though text analysis to see if it can predict smoking habits.

An interesting strategy to limit this problem is the one proposed by Collins et al. 1987. The authors collected saliva samples to increase the reliability of the self-declared smoking status. This test not only provides an objective measure to evaluate if an individual has smoked in the previous 24 hours, but it also proved to increase the amount of self-reported cigarettes consumption that an actual smoker reveals (Bauman, Koch, and Bryan 1982). This last point is particularly important for future research as it implies that the awareness of being tested increases the probability that a respondent will say the truth on his tobacco use. Therefore, it may prove sufficient to inform an individual that he will be tested to increase the reliability of his self-declared smoking habit, without the need of actually testing it.

Methodology Concerning the methodology adopted, our approach could be improved by modifying the cross-validation process applied for tuning the machine learning models. The best performing model was selected using the criterion of the highest classification accuracy. However, this measure is not the most appropriate way for evaluating the quality of a model. An alternative approach is to use the area under the curve (AUC) to assess the reliability of the predictions produced by a model. In this way we could tune in a more effective manner the machine learning model that we have implemented and achieve a stronger predictive power. However, it should be verified the computational feasibility of this approach, as computing the AUC for hundreds of models could prove extremely time consuming.

Another path that may lead to an increase in the accuracy of predictions is to implement different ML techniques than the ones adopted in this work. Given the binary nature of the dependent variable, support vector machine is clearly a valid option. Neural networks should also be considered.

Regarding the comparison with standard econometrics tools, we limited our analysis to OLS, but we could expand this comparison to non-linear parametric approaches.

Scope Finally, we need to keep in mind two characteristics of the presented work that limit its scope. The first is connected to the fact that in this analysis we want to identify a group of individuals who are the most likely to start smoking during adolescence (or at least a group of individuals with a higher probability than the average to start to smoke). In other words, we are focusing on the question: "Who are the individuals to whom it makes more sense to target an intervention to prevent smoking onset?". The answer is that these individuals are the ones identified by our model. However, we are not giving any guidance about the type of intervention

and its efficacy. Therefore, we are determining the group of individuals on which we want to intervene, but we are not giving any suggestion on how to do it.
A second important aspect is the one linked with the social implications of adopting this approach in a real policy problem. The main concern is that some parents would not accept that their child is assigned to a certain program based on an algorithm. The relevance of this issue should be verified in case the approach proposed would be put into practice.

2.6 Final remarks

In this chapter we developed different models to predict the smoking status of an adolescent based on his characteristics at age 10. When focusing on the 10% respondents considered most at risk of starting to smoke, the ensemble model could identify a group of adolescents with a percentage of actual smokers equals to 49%, against an average of the sample of around 16%. This method performed significantly better than all the others considered with the only exception of boosting.
While looking at the Area Under the Curve (AUC) no model could perform significantly better than another. Nonetheless, on the hold-out sample, ensemble method achieved an AUC of 73.69%.
These results are promising. Being able to identify a group of children that have a probability 3 times higher than the average to start to smoke during adolescence, could help policy makers to better design a prevention policy intervention. Moreover, we proposed various ways in which our study could be improved, fully exploiting the potential of machine learning. We believe that in future studies, adopting the correction proposed to our model, a relevant improvement in the predictive power of adolescents' smoking habits could be achieved.

Chapter 3

Application 2: Estimating HTE using causal forest

3.1 Background

In this chapter we investigate how machine learning algorithms can be adapted to estimate heterogeneous treatment effects (HTE). We focus on the so-called causal forest, proposed in Wager and Athey 2017[1]. After reviewing the theoretical assumptions on which this method is based, we propose an implementation on the Tennessee STAR Project, that aims at assessing the impact of small class sizes on students' performance in standardized tests.

Investigating heterogeneous treatment effects is crucial in different fields of applied research. In medicine for example, we want to know what are the patients that respond positively to a particular therapy, and those to which this therapy is not effective or even harmful. Each patient is treated in the same way, but, because of some individual characteristics, the response to the therapy is different. Finding out what are the features that makes the treatment effective or not, would improve the efficiency with which this therapy is given to patients.

In the same way, in economics, we are often interested in knowing the individuals that would benefit the most from a specific policy intervention or that are particularly affected by a treatment.

The traditional approach to study HTE in a linear regression consists in evaluating the significance of the coefficient of an interaction term between the treatment and a variable that we think may affect the treatment effect (TE). Alternatively, it is possible to sort the sample depending on a certain characteristic of interest and perform separate estimations on the subgroups obtained. In this way we can evaluate if the treatment effect differs in these subgroups.

[1] A growing literature developed other numerous approaches to study HTE with the use of machine learning (for instance Green and Kern 2012 and Hill and Su 2013).

Unfortunately, these approaches have important weaknesses. The most relevant is multiple testing: if we want to test heterogeneity with respect to many variables, we need to introduce numerous interaction terms, compromising the reliability of the coefficients' significance. Moreover, as pointed out by Wager and Athey 2017, we could be worried that a researcher would iteratively test the TE for different subgroups and then report only the most significants. Finally, using linear regressions we can study only pre-specified and linear relations.

Implementing causal forest, we can overcome these problems by obtaining individual treatment effects that are both asymptotically unbiased and Gaussian[2] (Wager and Athey 2017, Athey and G. Imbens 2016, Athey, J. Tibshirani, and Wager 2016). However, the only empirical implementation of this method that we could find in the literature is in Davis and Heller 2017b[3], where the authors evaluated the HTE of a summer job program that provided disadvantaged youth with a part-time job and mentoring.

We aim in this chapter at applying causal forest to the Tennessee STAR Project data, to study the HTE of assigning children to a class of small size on their performance in standardized tests. We start our analysis by outlining below the theoretical basis of causal forest.

3.2 Model set up

In order to produce asymptotically unbiased and Gaussian estimates with causal forest, Wager and Athey 2017 postulate the respect of various conditions. Following the potential outcomes framework (Rubin 1974), they define the treatment effect as follow:

$$\tau(x) = \mathbf{E}[y_i^{(1)} - y_i^{(0)} | X_i = x] \tag{3.1}$$

Where $y_i^{(1)}$ and $y_i^{(0)}$ represent the potential outcomes of an individual[4]. While, X_i is a vector of variables that we know are not affected by the outcome.

In order to estimate $\tau(x)$ we need to assume unconfoundedness, i.e. the potential outcomes $y_i^{(1)}$ and $y_i^{(0)}$ are independent of the treatment assignment W_i given the vector of features X_i[5]. In formula:

$$\{y_i^{(1)} \; y_i^{(0)}\} \perp W_i | X_i \tag{3.2}$$

[2]By individual treatment effect, we mean a treatment effect that is specific of each individual in the sample. Obtaining individual treatment effects allows us to perform a deeper analysis of HTE than traditional approaches.

[3]See also Davis and Heller 2017a.

[4]By potential outcomes we refer to the outcomes that we would obtain if we could observe the same individual with and without the treatment.

[5]It is very hard in observational studies to show convincingly that this condition is respected. However, as we will work with a treatment randomly assigned, we are sure that this condition is respected.

Moreover, the observations need to respect the stable unit treatment value assumption (Rubin 1978) and the conditional mean functions ($\mathbf{E}[y_i^{(1)}]$ and $\mathbf{E}[y_i^{(0)}]$) must be continuous. Finally, we need the following overlapping condition to be respected:

$$\epsilon < P[W = 1|X = x] < 1 - \epsilon \tag{3.3}$$

Where $\epsilon > 0$.

This last condition guarantees that if we have a sample with a large enough number of observations, then there will be sufficient control and treatment units in each leave of the tree to compute the conditional average treatment effect (CATE).

Respecting all the above-mentioned assumptions, Wager and Athey 2017 modify random forest in such a way that it produces asymptotically unbiased and Gaussian estimates of individual treatment effects. The aspects in which a causal forest differs from a random forest are two: the estimation process and the splitting algorithm.

Estimation process Concerning the first point, the random forest algorithm works by partitioning the features space and producing its estimates on the same sample. On the contrary, when running causal forest, the algorithm adopts a so-called "honest approach". It means that it randomly divides the initial sample in two subsamples, it uses the first one, the training sample (S^{Tr}), to create a partition of the features space and the second, the estimation sample (S^{Est}), to produce the estimates. Therefore, the partitioning and estimation processes are performed on two independent samples.

Splitting algorithm The second aspect in which random forest and causal forest differs is the splitting algorithm that it used to grow each tree that composes the forest.

The new splitting criterion is related to the maximization of the following formula:

$$-\widehat{MSE}_\tau(S^{Tr}, S^{Tr}, \Pi) = \frac{1}{N^{Tr}} \sum_{i \in S^{Tr}} \hat{\tau}^2(X_i; S^{Tr}, \Pi) \tag{3.4}$$

Where Π corresponds to a partitioning of the feature space, N^{Tr} is the number of observations in S^{Tr}, and $\hat{\tau}(.)$ is defined by the following formula:

$$\hat{\tau}(x; S, \Pi) = \hat{\mu}(1, x; S, \Pi) - \hat{\mu}(0, x; S, \Pi) \tag{3.5}$$

Where $\hat{\mu}(w, x; S, \Pi)$ represents the average of the observed y_i of those observation with $W_i = w$ and $X_i = x$, obtained using the subsample S and the feature space partition Π.

3.3 Data

Our analysis is based on the Tennessee STAR Project dataset. This study collects information on children in Tennessee that are attending kindergarten in the 1985-1986 school year. The children involved in this project are followed until the third grade. However, we will focus only on the data at kindergarten level.

The objective of the STAR Project is to evaluate the impact of belonging to a class of small size on the students' performance in standardized tests. In order to achieve this purpose, children are randomly assigned to one of three class typologies, before the beginning of the school year[6]. The three class typologies are: small classes, composed by a number of student between 13 and 17, regular classes composed by a number of student between 22 and 25, and regular/aide classes that have the same composition of a regular class but with the addition of a full-time teacher aide. The distribution of class typologies in the sample is reported in the following table[7]:

Table 3.1: Class typology distribution in the sample

Class typology	**Percentage**
SMALL	30.29%
REG/AIDE	35.09%
REGULAR	34.71%
Observations	5665

Between the end of March and the beginning of April, the students are tested to evaluate their math, writing and word recognition skills. Specifically, the test used to evaluate children's skills is the Stanford Achievement Test (SAT). The results of this tests constitute the dependent variable used to evaluate the impact of small class size on students' educational outcome.

We adopt the specification of the dependent variable proposed by Krueger 1999. This specification is realized by converting the test scores achieved in each section (math, writing, word recognition) to the corresponding percentiles. Subsequently, the percentiles obtained in the tree tests are averaged out to produce a measure of the overall performance of a student. This average represents the specification of the dependent variable that is used in in this study. Concerning the other variables available in the dataset that are used as controls, and that represent possible sources of heterogeneity in treatment effect, we provide descriptive statistics in table 3.2.

[6]This random assignment is particularly relevant for our analysis because one of the hypotheses on which causal forest is based is unconfoundness. Therefore, we are sure that this condition is respected thanks to random assignment.

[7]Observations with missing data are dropped from the sample. The total number of observations dropped is 660.

3.4 Previous results

Our analysis extends the results obtained by Krueger 1999, by exploring the drivers of heterogeneity in treatment effect with the use of causal forest. In this section we outline Krueger's findings that are relevant for our study.

Table 3.2: Descriptive statistics STAR Project

Variable	**Mean**
SAT percentile	49.15
Age month	68.60
Female	48.72%
Nonwhite	32.83%
Free lunch	48.29%
Teacher black	15.81%
Teacher experience	9.32
Special education	3.21%
Master or more	35.60%
Observations	5665

The table shows the descriptive statistics of the sample of the STAR Project considered in our analysis. Free lunch is a dummy that takes value one if a child receives free meals from the school.This variable is used as an indicator of social economic condition. Teacher experience is expressed in years. Master or more indicates if the teacher has obtained a master's degree or any postgraduate degree.

The author begins his analysis by estimating the following model:

$$Y_{ics} = \beta_0 + \beta_1 SMALL_{cs} + \beta_2 REG/AIDE_{cs} + \beta_3 X_{ics} + \alpha_s + \epsilon_{ics} \tag{3.6}$$

Where Y_{ics} is the SAT score, specified as described in the previous section, of student i, in class c at school s. $SMALL_{cs}$ and $REG/AIDE_{cs}$ represent dummy variables indicating respectively if a student was enrolled in a small class or in a regular class with an aide teacher. X_{ics} is a vector containing student and teacher's characteristics. α_s is a vector of dummies, one for each school involved in the program, that take value one when a student is enrolled in that specific school. In this way, school fixed effects are captured. Including these dummy variables is particularly important as random assignment was carried out within schools and not between schools. Finally, ϵ_{ics} is an error term.

The results of the implementation of this model on the STAR project data, show that attending a small class increases the average performance in kindergarten of about 5 percentile points with respect to the regular size class[8]. On the contrary, the effect of the aide teacher in the regular size class is not significant[9].

Concerning the explanatory power of control variables, we observe a significant effect of ethnicity, that is specified with a dummy variable that takes value one if the student is white or Asian and has a positive effect, sex, as girls perform significantly better than boys, and of socio-economic conditions, that are measured with a dummy variable that takes value one when a child receives free lunch. As expected, this last variable has a strong negative effect on SAT performance. On the contrary, teacher's characteristics (experience, education, sex) do not strongly impact the students' results on the SAT. Only teacher experience has a significant impact on the dependent variable. Each year of marginal experience increases on average the SAT performance of 0.26 percentile points.

Given that the different classes typologies had different sizes within each category, Kreuger also implement a different specification of the model in which the variable of interest is now the actual class size instead of a dummy that represents the class typology. As actual class size is not randomly assigned, the author instrumentalizes it with class typology and implement a 2SLS model. The estimates obtained with this different specification are consistent with the previous ones, as an increase of one student in the class is associated with a decrease of 0.71 in the percentile SAT results of the students of that class[10].

Finally, the most relevant part of Krueger's article for our purpose is the one in which he investigates the heterogeneity of treatment effects. In order to carry out this kind of analysis the author divides the sample in different subgroups and studies the treatment effect on these subgroups[11]. The sample is sorted by sex, ethnicity, socio-economic conditions and region of residence (rural, metropolitan, inner-city, towns). The author finds that smaller classes have a larger effect on boys, students on free lunch, non-white and living in the inner-city. It seems therefore that those children who on average perform worse could benefit the most from the assignment to a small class.

These results provide useful guidance for future policy application as they single out the groups who would benefit the most from the treatment. However, these results are biased toward a previous conception of the phenomenon, because the subgroups are formed following what the

[8]This result is significant at any confidence level.

[9]The author suspects that the availability of aide teacher could confound the true effect of teaching aid.

[10]Again this result is significant at any confidence level.

[11]This analysis is performed on a pooled OLS model, where the author gathers data from all the four years of the study available (from kindergarten to third grade). The results outlined in this section refers only to kindergarten level assignment.

author thinks will impact the treatment effect. The strength of causal forest is that we can perform the same analysis without restricting the features considered as possible sources of heterogeneity in treatment effect[12].

3.5 Econometric strategy

In this section we present the econometric specification of the model that we adopt in our analysis and our strategy to make use of the individual treatment effects produced by causal forest.

As we already pointed out, the dependent variable is coded in the same way as Krueger 1999, using the percentiles average in the three sections as the measure of students' achievement. Also the controls included in the model are mostly the same, with only marginal differences in the way in which they are specified. Moreover, we do not include teacher's sex, given that the male percentage is extremely low, but we include a dummy that takes value one if a student receives special aid, a variable that was not considered in Krueger's regressions.

As a first step, we simply replicate the OLS model performed by Krueger with our slightly different specification, obtaining obviously extremely similar results, that are presented in the next section.

Subsequently, we run random forest on this model to study the heterogeneity in treatment effects. In this way, we find the individual treatment effects and the associated confidence intervals. We rank observations by individual treatment effects and understand if there is a significant difference between the individuals with the lowest TE and the ones with the highest one. Moreover, we compute for what portion of the sample the TE is significantly different from zero.

Furthermore, we study what are the variables that drives the heterogeneity in treatment effects and plot the functional form of the relations between a certain feature and the treatment effect.

Finally, we create two subgroups, the quarter estimated to have the lowest TE (hereinafter referred to as Q1) and the quarter estimated to have the highest (hereinafter referred to as Q4), and compare their characteristics and test if their average treatment effects are actually different. Performing this last analysis give us a picture of how a policy intervention of this kind could be improved in terms of efficiency by targeting a subgroup identified via causal forest.

[12]It is possible to perform the same analysis also with the strategy proposed by the author, by dividing the sample in hundreds of different ways depending on all the characteristics. However, this approach would suffer from the weaknesses previously explained, and it would be unfeasible with a high number of features.

3.6 Results

OLS We start to present the results obtained in our analysis with the replication of Krueger's OLS regression, using our slightly different specification.

These results are showed in table 3.3. Belonging to a small class increases on average the performance in the SAT of 4.68 percentile points with respect to the regular class. On the contrary, belonging to the regular class with teacher aide does not produces any significant impact. These estimates are consistent with Krueger's and very close in magnitude. The small variations in the coefficients are due to the slightly different specification of the control variables and the inclusion of the dummy that captures if a student receives special aid. Age has a significant impact on SAT achievements (the older the child, the higher the percentile scored). Moreover, white children score almost 12 percentile points higher than non-white children and students who receive free lunch around 13 points lower than those who do not. Finally, girls score 4 points better than boys.

Concerning teachers' characteristics, the years of experience represent the only feature that impacts significantly student's performance. Each year of marginal experience produces on average an increase of 0.25 percentile points in the SAT.

Causal forest Now we focus our attention on the study of HTE using causal forest. First of all, we compute the individual treatment effects with this method and rank them in ascending order. In figure 3.1 the individual TEs are plotted with and without confidence intervals. The left graph of figure 3.1 is showing a remarkable heterogeneity in the estimated individual treatment effects. Indeed, some of the children have an estimated TE that is negative, while others have an estimated TE of above 15 percentile points.

The right graph of figure 3.1, where we added confidence intervals, allows us to draw two relevant conclusions. The first, is that the observations at one extremity of the plot are significantly different in estimated TE from the ones at the other extremity. Therefore, heterogeneity in treatment effect is statistically significant. The second, is that about half of the considered individuals do not have a treatment effect that is significantly higher than zero. This means that only for the other half of the sample we are sufficiently sure that the small class size is improving learning outcome. This point is relevant for our analysis, as a policy maker is interested in providing a costly service, smaller classes, only to those individuals to whom providing this service has an effect that is significantly different from zero.

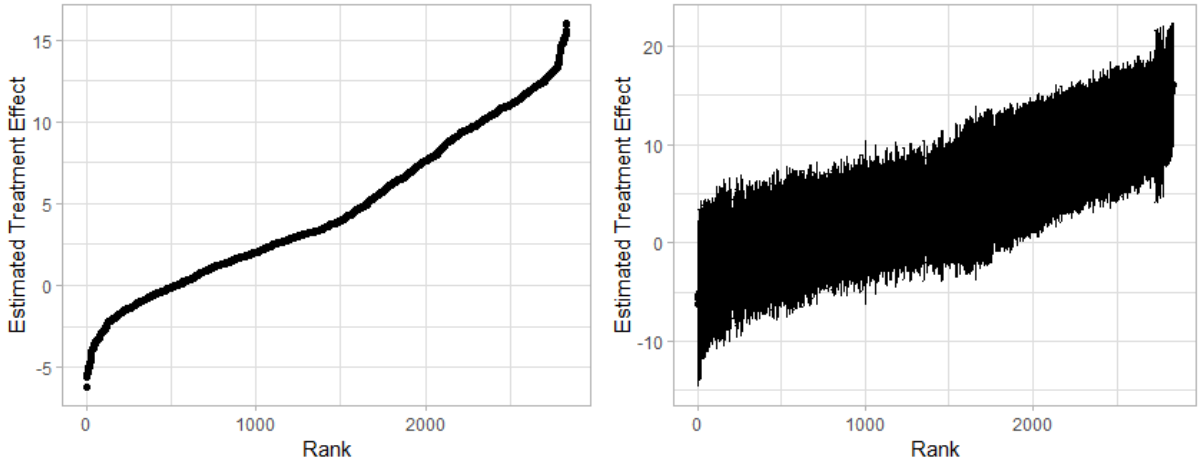
Table 3.3: OLS regression estimates

VARIABLES	SAT PERCENTILE
SMALL	4.677***
	(0.702)
REG/AIDE	-0.274
	(0.678)
Age month	0.636***
	(0.068)
Female	4.054***
	(0.559)
Nonwhite	-11.79***
	(1.159)
Free lunch	-13.36***
	(0.677)
Teacher black	1.905*
	(1.119)
Teacher experience	0.247***
	(0.055)
Special education	-12.80***
	(1.661)
Master or more	-0.338
	(0.734)
Constant	-14.64**
	(5.875)
School fixed effects	Yes
Observations	5,665
R-squared	0.344

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Figure 3.1: Individual treatment effects with and without confidence intervals



The graph on the left plots the individual treatment effects computed with causal forest ranked from the lowest to the highest. In the plot on the right we added confidence intervals.

To further investigate heterogeneity, we divide our sample in quarters, based on the estimated treatment effect, and study the average treatment effect for these subgroups. Specifically, we consider the one fourth of the individuals with the highest estimated TE (Q4) and the one fourth of the individuals with the lowest TE (Q1) and compare the average treatment effects of these two subgroups. In order to do it, we run the same OLS regression that we ran on the entire sample on the two subgroups. In table 3.4, we report the results of these regressions.

The effect of belonging to a small class on SAT achievement in the subgroup with the lowest estimated TE is insignificant. While in Q4, this effect is quantified in around 11 percentile points of increase, and it is significant at any confidence level and significantly different from the effect obtained on the whole sample[13]. These regressions confirm that causal forest correctly singled out individuals with particularly high and low treatment effects.

Interestingly, in the first subgroup, the effect of being in a regular class with a teacher aide has a significant negative impact. Even if this result should be taken carefully, given the reduced dimension of the sample considered, it is consistent with the fact that in this subgroup the average treatment effect of a small class is not different from zero. Indeed, we could have identified a group of children that perform better when they are not in strict contact with a teacher. For this reason, when included in a smaller class or in a class with an aide teacher, their performance does not improve, or it even worsen.

In addition to looking at the intensity of treatment effect in the two subgroups, we focus here on their descriptive statistics, in order to understand what are the characteristics that are correlated

[13] We should also consider that the sample dimension is now strongly reduced, therefore significance at any level is not trivial given this fact.

with a higher or lower TE. This analysis is particularly relevant from a policy making point of view, as these characteristics could represent guidelines to target a policy intervention to those individuals who would benefit the most from it. Table 3.5 reports the descriptive statistics of the two groups and the level of significance of the differences between them.

The table shows that the two subgroups are significantly different in all the characteristics considered. Individuals in Q1 are on average about one month older than those in Q4. Moreover, the percentage of female is much lower in the first subgroup than in the second. Concerning ethnicity and socio economic conditions, the children who benefit the most from the treatment are on average poorer (as 68.50% of them receive free lunch, against a 27.01% in Q1) and non-white (50.56% against 14.00%) with respect to Q1. Concerning teacher's characteristics, the children in Q4 have on average a teacher with lower experience, who is more frequently black and with a master's degree.

We conclude our analysis of heterogeneous treatment effects focusing on the functional form of the relation between a specific feature and the treatment effect. Figure 3.2 plots the estimated treatment effect as a function of the years of experience of a teacher. The blue line represents the conditional (on a determined level of experience) average treatment effect. Consistently with the results obtained by studying the characteristics of Q1 and Q4, we observe that the TE is stronger for those teachers with less experience. In particular, the estimated TE peaks at 5 and 6 years of experience. On the contrary, for those teachers with more than 10 years of experience the treatment effect is close to zero. In general, we observe a remarkable heterogeneity in TE with respect to the years of teaching experience.

Figure 3.3 plots the estimated treatment effect as a function of the children age. If we exclude the two extremities of the graph, that are affected by the low number of observations, we do not notice any significant pattern. Therefore, we can conclude that children age do not strongly impact the TE [14].

3.7 Final remarks

In this chapter we implemented causal forest to study heterogeneity in treatment effects in the Tennessee STAR project. Adopting causal forest enabled us to study more in depth the drivers of heterogeneity and avoid the weaknesses of traditional approaches to this topic.

The aim of this analysis was to provide useful information on how a policy intervention should be designed. Indeed, knowing the characteristics of the children that are the most affected by the treatment could help policy makers in shaping targeted and effective policies.

[14] It is to remark that variability in age is very low. This fact could affect our results.

We found out that the treatment (belonging to a class of small size), produces significantly heterogenous effects that depend on various characteristics of the children involved in the study. More specifically, the children who benefited the most from the treatment are non-white, male and poorer then the children who benefited the least. These results are consistent with Krueger's findings.

Concerning teachers' characteristics, we determined that a high TE is correlated with lower level of experience. In particular, the highest level of TE are obtained for those teachers with around five years of experience. On the contrary, teachers with an experience of more than 10 years are correlated with close-to-zero treatment effect.

Table 3.4: OLS regressions on Q1 and Q4

VARIABLES	SAT PERCENTILE Q1	SAT PERCENTILE Q4
SMALL	0.028	11.03***
	(2.644)	(2.402)
REG/AIDE	-5.949**	1.900
	(2.488)	(2.501)
Age month	0.751***	0.341
	(0.196)	(0.226)
Female	4.054***	8.173***
	(1.769)	(2.115)
Nonwhite	-10.47*	-14.22***
	(5.808)	(3.530)
Free lunch	-18.75***	-15.07***
	(2.339)	(2.406)
Teacher black	-26.12**	7.986**
	(10.34)	(3.099)
Teacher experience	0.423**	0.339
	(0.198)	(0.564)
Special education	-5.690	-14.73**
	(4.090)	(6.417)
Master more	-7.346**	0.890
	(3.204)	(2.959)
Constant	1.000	14.33
	(18.813)	(23.44)
School fixed effects	Yes	Yes
Observations	708	708
R-squared	0.372	0.421

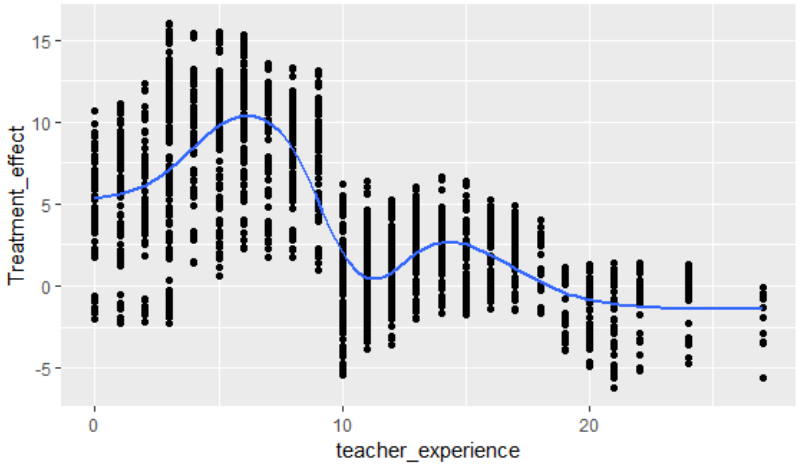
Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 3.5: Comparison between Q1 and Q4

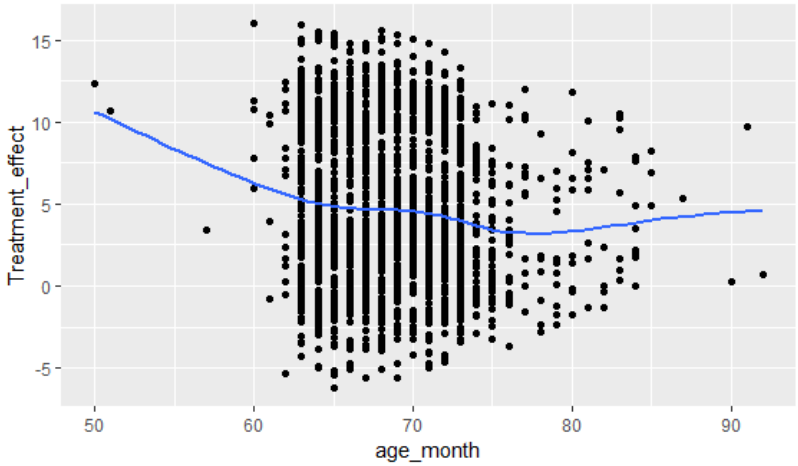
Variable	Q4 mean (a)	Q1 mean (b)	Significance a-b
Age month	67.72	68.65	* * *
Female	29.66%	57.85%	* * *
Nonwhite	50.56%	14.00%	* * *
Free lunch	68.50%	27.01%	* * *
Teacher black	23.73%	5.94%	* * *
Teacher experience	5.639	13.879	* * *
Special education	1.18%	4.38%	* * *
Master more	37.42%	27.29%	* * *

Figure 3.2: Treatment effects as a function of teacher's years of experience



The graph plots individual treatment effects as a function of teacher's experience expressed in years. The blue line represents the conditional (on teacher's experience) average treatment.

Figure 3.3: Treatment effects as a function of students' age



The graph plots individual treatment effects as a function of students' age expressed in months. The blue line represents the conditional (on age) average treatment effect.

Conclusions

In this thesis we presented two applications of machine learning (ML) techniques as an instrument for policy making and policy impact evaluation.

In the first one, we focused on a predictive problem: identifying a group of adolescents with a high probability of starting to smoke in the near future. Using the British Cohort Study dataset, we exploited the individual characteristics at age 10 to predict the smoking status at age 16. We could identify a group, the 10% with the highest predicted probability of starting to smoke, where the percentage of actual smokers is 49%. This means more than 3 times the percentage of smokers in the entire sample, that is 16%. Being able to identify this group, provides useful guidance to policy makers for targeting the individuals who need the most an intervention, for example through a mentoring activity that increases the awareness of the risks connected to smoking.

Machine learning performed significantly better than OLS in identifying this group. The ensemble method, that averaged out the predicted probabilities of all the ML models implemented, proved to produce the most accurate predictions. However, while adopting the area under the ROC curves (AUC) as the criterion used to assess predictions' quality, no method could perform significantly better than another. Nonetheless, in the hold-out sample, the ensemble model produced the highest AUC, equal to around 73%.

We suggested different ways in which our approach could be improved. We expect that by adopting some of the recommendations proposed, the prediction accuracy of adolescents' smoking status could be increased. However, this work represents a starting point for future research as the methodology implemented can be replicated and the results achieved used as a benchmark for future studies.

In the second application of this thesis, we focused on a fundamentally different problem. We exploited ML techniques to study heterogeneity in treatment effects (HTE), adopting the so-called causal forest approach. Specifically, we studied the effect on standardized tests of attending kindergarten in a class of small size, compared to a class of regular size. To produce this analysis, we used the data from the Tennessee STAR Project. Thanks to the implemen-

tation of causal forest we could obtain individual treatment effects and the related confidence intervals. This allowed us to examine HTE in a deeper way than previous studies. We found that the treatment produced significantly heterogeneous effects. In particular, the children who benefited the most from the treatment are non-white, male and poorer then the children who benefited the least. Moreover, teachers' experience proved to be a relevant driver of heterogeneity. Teachers with about 5 years of experience are correlated with the highest TE. On the contrary, those with an experience of more than 10 years are correlated with close-to-zero treatment effect.

Bibliography

Acemoglu, Daron, Simon Johnson, and James A Robinson (2001). "The colonial origins of comparative development: An empirical investigation". In: *American economic review* 91.5, pp. 1369–1401.

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

Athey, Susan and Guido W Imbens (2017). "The state of applied econometrics: Causality and policy evaluation". In: *Journal of Economic Perspectives* 31.2, pp. 3–32.

Athey, Susan, Julie Tibshirani, and Stefan Wager (2016). "Generalized random forests". In: *arXiv preprint arXiv:1610.01271*.

Bauman, Karl E, Gary G Koch, and Elizabeth S Bryan (1982). "Validity of self-reports of adolescent cigarette smoking". In: *International Journal of the Addictions* 17.7, pp. 1131–1136.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014). "Inference on treatment effects after selection among high-dimensional controls". In: *The Review of Economic Studies* 81.2, pp. 608–650.

Breiman, Leo (1996). "Bagging predictors". In: *Machine learning* 24.2, pp. 123–140.

— (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

— (2017). *Classification and regression trees*. Routledge.

Chandler, Dana, Steven D. Levitt, and John A. List (2011). "Predicting and Preventing Shootings among At-Risk Youth". In: *American Economic Review* 101.3, pp. 288–92.

Chen, Jiajian, Wayne J Millar, et al. (1998). "Age of smoking initiation: implications for quitting". In: *Health reports-statistics Canada* 9, pp. 39–48.

Chung, Sophia and Youngji Li (2017). "An Application of Machine Learning for the Identification of Adolescent Smoking Risk Factors". In:

Collins, Linda M et al. (1987). "Psychosocial Predictors of Young Adolescent Cigarette Smoking: A Sixteen-Month, Three-Wave Longitudinal Study 1". In: *Journal of Applied Social Psychology* 17.6, pp. 554–573.

Cortez, Paulo and Alice Maria Gonçalves Silva (2008). “Using data mining to predict secondary school student performance”. In:

Davis, Jonathan and Sara B Heller (2017a). *Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs*. Tech. rep. National Bureau of Economic Research.

— (2017b). “Using Causal Forests to Predict Treatment Heterogeneity: An Application to Summer Jobs”. In: *American Economic Review* 107.5, pp. 546–50.

Dumortier, Antoine et al. (2016). “Classifying smoking urges via machine learning”. In: *Computer methods and programs in biomedicine* 137, pp. 203–213.

Fawcett, Tom (2006). “An introduction to ROC analysis”. In: *Pattern recognition letters* 27.8, pp. 861–874.

Frank, Charles, Asmail Habach, and Raed Seetan (2018). “Predicting smoking status using machine learning algorithms and statistical analysis”. In: *Journal of Computing Sciences in Colleges* 33.3, pp. 66–66.

Green, Donald P and Holger L Kern (2012). “Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees”. In: *Public opinion quarterly* 76.3, pp. 491–511.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). “Unsupervised learning”. In: *The elements of statistical learning*. Springer, pp. 485–585.

Hill, Jennifer and Yu-Sung Su (2013). “Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children’s cognitive outcomes”. In: *The Annals of Applied Statistics*, pp. 1386–1420.

James, Gareth et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Kleinberg, Jon et al. (2015). “Prediction Policy Problems”. In: *American Economic Review* 105.5, pp. 491–95.

Krueger, Alan B (1999). “Experimental estimates of education production functions”. In: *The quarterly journal of economics* 114.2, pp. 497–532.

Kuhn, Max et al. (2008). “Building predictive models in R using the caret package”. In: *Journal of statistical software* 28.5, pp. 1–26.

Lantz, Brett (2015). *Machine learning with R*. Packt Publishing Ltd.

McCaffrey, Daniel, Greg Ridgeway, and Andrew Morral (2004). “Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies”. In: *Psychological Methods* 4.9, pp. 403–425.

Mullainathan, Sendhil and Jann Spiess (2017). “Machine Learning: An Applied Econometric Approach”. In: *Journal of Economic Perspectives* 31.2, pp. 87–106.

Rubin, Donald B (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of educational Psychology* 66.5, p. 688.

— (1978). “Bayesian inference for causal effects: The role of randomization”. In: *The Annals of statistics*, pp. 34–58.

Saddleson, ML et al. (2016). “Assessing 30-day quantity-frequency of US adolescent cigarette smoking as a predictor of adult smoking 14 years later”. In: *Drug and alcohol dependence* 162, pp. 92–98.

Suchting, Robert et al. (2017). “Using elastic net penalized cox proportional hazards regression to identify predictors of imminent smoking lapse”. In: *Nicotine & Tobacco Research*.

Thomas, Roger E et al. (2015). “Family-based programmes for preventing smoking by children and adolescents”. In: *Cochrane Database of Systematic Reviews* 2.

Tibshirani, Robert (1996). “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288.

Wager, Stefan and Susan Athey (2017). “Estimation and inference of heterogeneous treatment effects using random forests”. In: *Journal of the American Statistical Association* just-accepted.

Wellman, Robert J et al. (2016). “Predictors of the onset of cigarette smoking: a systematic review of longitudinal population-based studies in youth”. In: *American Journal of Preventive Medicine* 51.5, pp. 767–778.

Appendices

Appendix A

Different thresholds

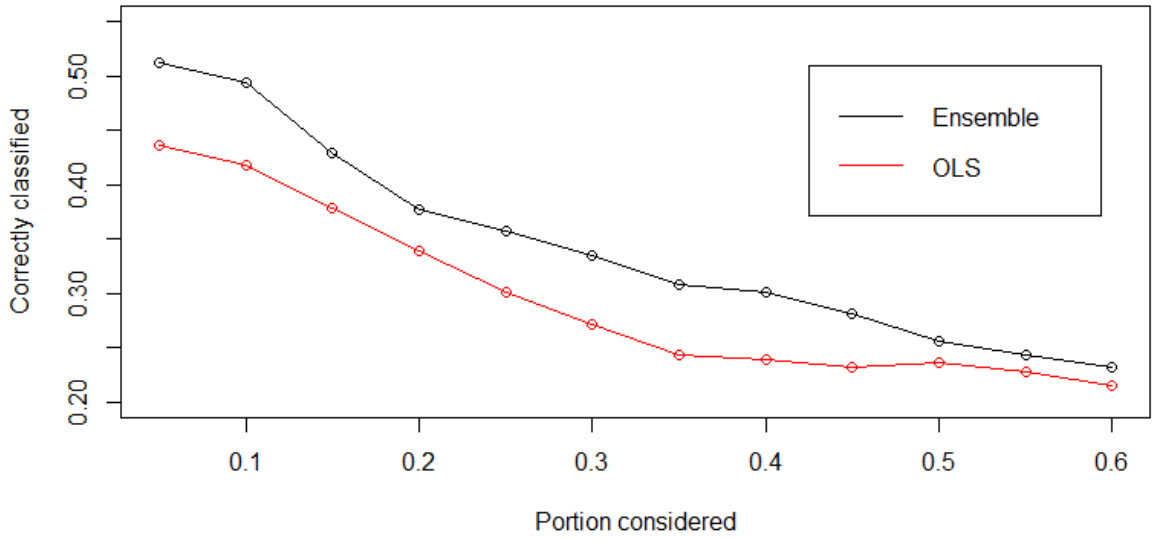
In our analysis we focused on the predictive power of the model on the individuals considered as the 10% most at risk of starting to smoke. First, we computed the probability of starting to smoke at age 16 with the different models. Subsequently, we sorted out the 10% individuals with the highest probability. Finally, we evaluated what percentage of this group actually started to smoke. This approach is particularly useful as it could be adopted by a policy maker to single out the individuals to whom he would target a policy intervention.

However, the choice of a 10% is clearly arbitrary. In this appendix, we evaluate how changing this threshold affects our results in terms of accuracy. In the graph below, we plotted the percentage of actual smokers that a model[1] can identify in the group of the x% individuals with the highest probability of starting to smoke. We considered the value of x from 5% to 60%, using 5% intervals.

The graph shows that by choosing a 5% threshold, the ensemble method can identify a group with a percentage of smokers around 52%, while OLS is around 44%. When a threshold higher than 10% is chosen, we note that the percentage of actual smokers drops quickly. For example, when it is 20%, implementing ensemble the percentage is lower than 40% and OLS lower than 35%. When we look at higher threshold the performance of OLS and ensemble converges, producing extremely similar results. This fact is also consistent with the results obtained using ROC curves, that were not significantly different between OLS and machine learning models. Nonetheless, even adopting a 60% threshold, these models identify a group with a percentage of actual smokers around 25%. This is still a remarkable result considered that in the whole sample the portion of smokers at age 16 is 16%.

[1]The models take into account in this appendix ae only ensemble, that proved to be the best performing model using a 10% threshold, and OLS that is mainly used as a comparison.

Figure A.1: Percentage of individuals most at risk correctly classified as a function of the threshold used



The graph plots how the percentage identified by a model as the x% most at risk that are actual smokers varies as x is moved from 5% to 60%

Appendix B

Alternative specifications

B.1 Basic demographics

In this appendix we present the results in terms of predictive power obtained including in the models only basic demographic information and those related to relatives and friends' smoking habits. The number of predictors considered is 52. However, a relevant number of these predictors is represented by dummy variables that specify qualitative variables [1].

The main scope of this appendix is to understand if making use of a high number of features produces better predictions. We will focus on the two most important criterion adopted in this thesis to evaluate predictions' quality: the AUC of the ROCs obtained implementing the different models, and the percentage of actual smokers in the 10% of individuals predicted to be most at risk of starting to smoke at age 16.

We start by looking at the results achieved in this last measure. The table below reports the percentage of actual smokers among the individuals identified as the 10% most at risk of starting to smoke. We compare the outcome achieved with the new specification and the one of the main analysis.

Table B.1 shows clearly that using only basic demographic information leads to a remarkable decrease in prediction accuracy of the individuals most at risk of staring to smoke. However, this decrease is particularly strong for machine learning methods but barely significant for OLS. This result is in line with our expectations as reducing the variables included in the model can favour OLS by reducing overfitting. On the contrary, it makes more difficult for machine learning methods to find relevant patterns.

Finally, we need to stress that a relevant part of the variable included in this new specification are dummies used to describe qualitative variables. This specification is particularly appropriate

[1] For instance, there are 10 variables concerning the region where the child lives.

for linear regression and could explain why OLS can perform better than some machine learning methods.

Table B.1: 10% most at risk performance comparison

Method used	**Percentage of actual smokers original specification**	**Percentage of actual smokers only basic info**
OLS	41.77%	37.62%
	[40.03%, 43.51%]	[34.48%, 39.78%]
Random Forest	42.85%	29.70%
	[40.72%, 44.99%]	[27.27%, 30.53%]
Boosting	48.10%	33.66%
	[45.83%, 50.37%]	[31.97%, 36.67%]
LASSO	43.04%	29.70%
	[41.44%, 44.63%]	[28.54%, 31.43%]
Ensemble	49.37%	34.65%
	[45.55%, 53.23 %]	[30.65%, 37.80 %]

The table reports the percentage of actual smokers in the group of individuals identified by the different models as the 10% most at risk of smoking, using two different specifications. The first consider all the variables used in our analysis. The second only basic demographic information and those on relatives and friends' smoking habits. In parenthesis the bootstrapped confidence intervals are reported.

Table B.2 shows the comparison between AUCs obtained with the two different specifications. The performance of OLS is almost identical in the two cases, confirming that the loss of information due to the lower number of variables included in the model is offset by a lower overfitting. On the contrary, we observe a relevant decrease in then AUC achieved in the hold-out-sample while implementing machine learning. This decrease is significant at a 5% level only for ensemble. However, it is significant at 10% level also for random forest and boosting, but not for LASSO.

Therefore, this second criterion to evaluate predictions' quality also provides evidences that making use of the whole set of variables is useful to improve it. However, this is true for most machine learning methods but not for OLS and LASSO.

Table B.2: Comparison between Area Under the Curves

Method used	**Area Under the Curve original specification**	**Area Under the Curve only basic info**
OLS	66.73%	66.75%
	[61.14%, 71.70%]	[62.25%, 71.24%]
Random Forest	72.62%	63.13%
	[67.46%, 77.57%]	[58.46%, 67.79%]
Boosting	73.02%	64.22%
	[67.75%, 77.41%]	[59.52%, 68.89%]
LASSO	69.61%	62.03%
	[63.95%, 74.39%]	[57.26%, 66.81%]
Ensemble	73.69%	63.76%
	[69.03%, 78.58%]	[59.07%, 68.46%]

The table shows the AUCs obtained with the two different specifications. In parenthesis the bootstrapped confidence intervals are reported.

B.2 Smoking status

In our main analysis, we have dropped from the sample the individuals who smoked between 1 and 4 cigarettes per week and less than a cigarette per week. We have focused in this way on a classification between regular smokers and non-smokers. In this section we explore a different specification of the dependent variable.

We included among the smokers the individuals who consumed between 1 and 4 cigarettes per week and among the non-smokers the individuals who consumed less than one cigarette per week.

This new specification did not produce a significant change in the AUC obtained by the different models. However, it leaded to a significant worsening on the portion of smokers in the group of the 10% most at risk. Indeed, no model could find a group with a percentage of actual smokers higher than 40%. This worsening is expected because we are considering as smokers also some individuals who make use of tobacco only occasionally, and as non-smokers individuals that sometimes do smoke. Therefore, it becomes harder for the models to correctly classify individuals, as the characteristics of the two groups are now more similar.

The results founded are summarized in the tables below.

Table B.3: 10% most at risk performance comparison

Method used	**Percentage of actual smokers original specification**	**Percentage of actual smokers alternative specification**
OLS	41.77%	35.36%
	[40.03%, 43.51%]	[33.33%, 39.18%]
Random Forest	42.85%	29.62%
	[40.72%, 44.99%]	[27.47%, 30.77%]
Boosting	48.10%	39.02%
	[45.83%, 50.37%]	[37.21%, 43.14%]
LASSO	43.04%	34.14%
	[41.44%, 44.63%]	[26.23%, 37.71%]
Ensemble	49.37%	37.80%
	[45.55%, 53.23 %]	[33.63%, 39.75 %]

The table reports the percentage of actual smokers in the group of individuals identified by the different models as the 10% most at risk of smoking, using two different specifications. In the original specification, Those individuals who declared to smoke between one and four cigarettes per week and less than one per week are dropped from the sample. In the alternative specification they are included respectively in the smokers and non-smokers groups. In parenthesis the bootstrapped confidence intervals are reported.

Table B.4: Comparison between Area Under the Curves

Method used	**Area Under the Curve original specification**	**Area Under the Curve alternative specification**
OLS	66.73%	66.15%
	[61.14%, 71.70%]	[61.06%, 70.81%]
Random Forest	72.62%	63.53%
	[67.46%, 77.57%]	[58.68%, 68.37%]
Boosting	73.02%	66.74%
	[67.75%, 77.41%]	[61.97%, 71.51%]
LASSO	69.61%	64.76%
	[63.95%, 74.39%]	[59.86%, 69.66%]
Ensemble	73.69%	66.67%
	[69.03%, 78.58%]	[61.88%, 71.47%]

The table shows the AUCs obtained with the two different specifications. In parenthesis the bootstrapped confidence intervals are reported.