Predicting Marketing Campaign's Conversion for Banking Institutions

15.095 – Machine Learning Under a Modern Optimization Lens

DECEMBER 9, 2022

Ahmad Hussain <ahmad5@mit.edu>
Nikolaos Galanos <nikosgal@mit.edu>

1 - Introduction

With the advent of digital banking, traditional banking channels like branches and ATMs are fast becoming obsolete, and customers are now more demanding and dissatisfied than ever before. The financial disruptions due to the COVID-19 pandemic created unique opportunities and challenges for economic inclusion, some of which may be temporary, while others may be longer lasting. According to a 2021 FDIC survey the use of new digital products offered by banks has inclined and this gives the opportunity to banks to compete not only on interest rates but on other offerings and bundles. To achieve the best results, the banks should carefully design their marketing campaigns and pick their target audience.

Our project focuses on an analysis of direct marketing campaigns for a Portuguese bank. Direct marketing campaigns form an important customer acquisition channel for traditional banks through which they seek to convert prospective customers into new clients. However, conversion rates for direct marketing campaigns in the banking industry remain low with estimates placing them under 5%. With the increased penetration of digital banking services, there are larger industry tailwinds that orient customers towards modern banking services. A well-designed direct marketing campaign is critical to conversion success as opposed to traditional approaches that adopt a one-size-fits-all approach.

2 - Problem Summary

Our core motive through this project is to underscore how optimization and machine learning can allow us to rethink traditional direct marketing and in the process realize greater economic returns. The classification techniques discussed are highly potent and practical; through this analysis, we aim to underline their power in applied settings. In terms of our aims, we seek to build classification models which predict whether a particular customer will make a term deposit with the bank. A successful understanding of this can allow banks to adopt leaner marketing campaigns with increased return on each dollar spent for returns. We combine classification models with interpretable clustering that allows us to segment the target customer base and understand more effectively the composition of target segments. By undertaking clustering on the dataset, we can develop a tangible understanding of the underlying patterns behind each segment. Finally, we utilize prescriptive techniques by slightly modifying our problem, in order to help the bank decide on the best medium to use to contact a customer in an attempt to maximize deposits.
Our problem could easily be generalized to other financial institutions and for other ways of approaching the customer. Given the goal of minimizing the resources in terms of money, time and personnel that each institution utilizes in their attempt to maximize profits, we believe that the problem of identifying the customers that are more prone to a new offering and targeting them is of crucial significance.

3 - Dataset

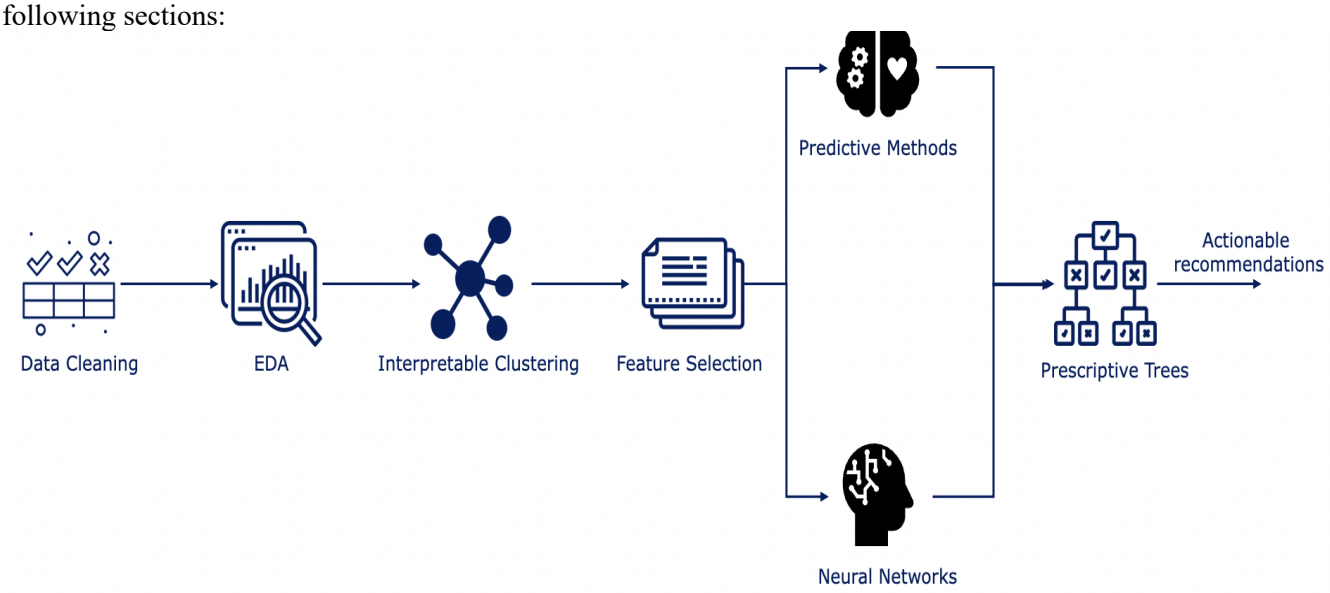
For our project we will use a dataset hosted at UCI Machine Learning Repository with data related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The dataset provided consists of 41,188 rows, with each row corresponding to a direct call to a customer. For each call we have 20 features related not only to customer and campaign characteristics but also with some market indexes such as the Euribor index and the Consumer Price Index. The variable (y) to predict is whether a customer will subscribe to a term deposit, making our problem a classification problem. Our features could be summarized in the following table:

Demographic	Financial	Campaign specific	Social & Employment
Age	Default history (y/n/unknown)	Past campaign target (y/n)	Employment Variation Rate Index
Occupation (12 values)	Housing loans(y/n/unknown)	Days since last contact	Consumer Price Index
Marital Status (4 values)	Personal loans(y/n/unknown)	Previous number of contacts	Consumer Confidence Index
Education level (8 values)		Outcome of previous contact(3 values)	Euribor 3 month Index
		Duration of call in seconds	Number of bank employees

		Communication type	
		Day/month of last contact	

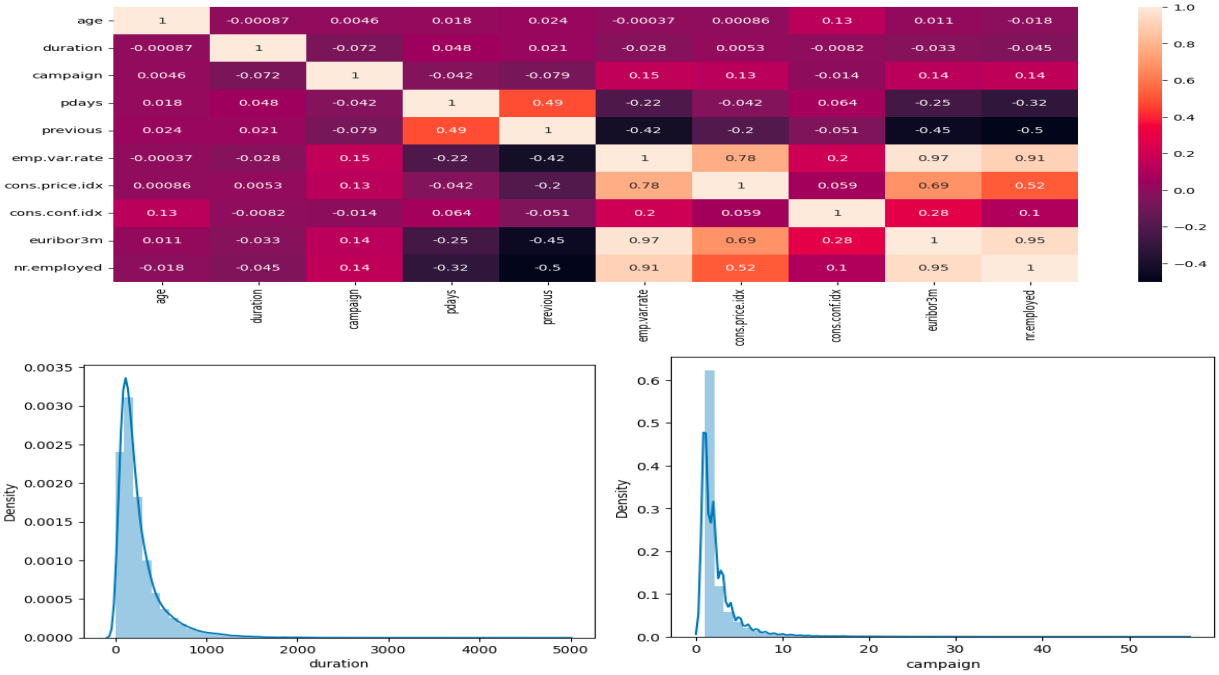
4 - Methodology

Our larger approach across the project can be broken down as follows: Each step is thoroughly described on the following sections:



5 - Exploratory Data Analysis

To get a better idea of our dataset and to understand relationships between our features as well as get an idea of the data preprocessing and feature engineering that we would have to do, we performed exhaustive Exploratory Data Analysis.



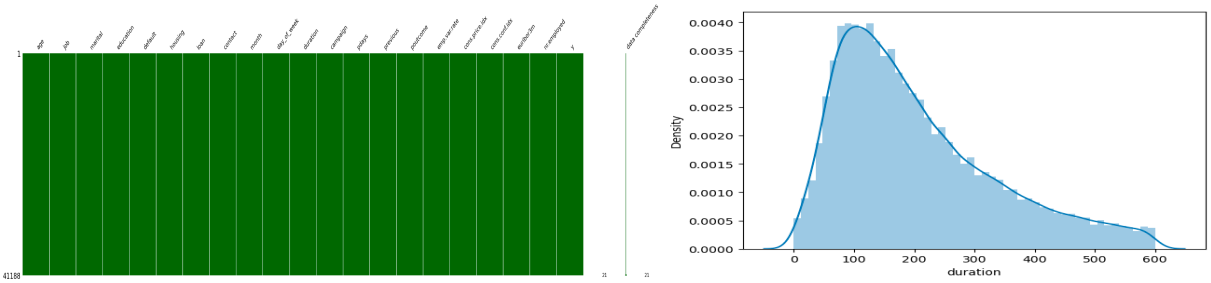
Some of our key findings include the fact that most of our variables (except the market indexes) are not highly correlated. The feature that tracks call duration seems to be highly correlated with the final outcome of the call (whether the customer made a deposit or not). Moreover, people with above average call durations seem to be the ones that make deposits; however, this

is a variable that we cannot know and control before a call, hence we will not use it as a predictor for our models. In addition, we notice that our dataset contains some outliers on the features duration and campaign.

In terms of more impactful business decision findings, our exploratory data analysis showed that Single people between 20 and 30 with education levels greater than high school are more prone in making a deposit. On the other hand, people who currently have a loan (personal, housing) are less prone to making a deposit. (Relevant Graphs in the Appendix)

6 - Data Preprocessing

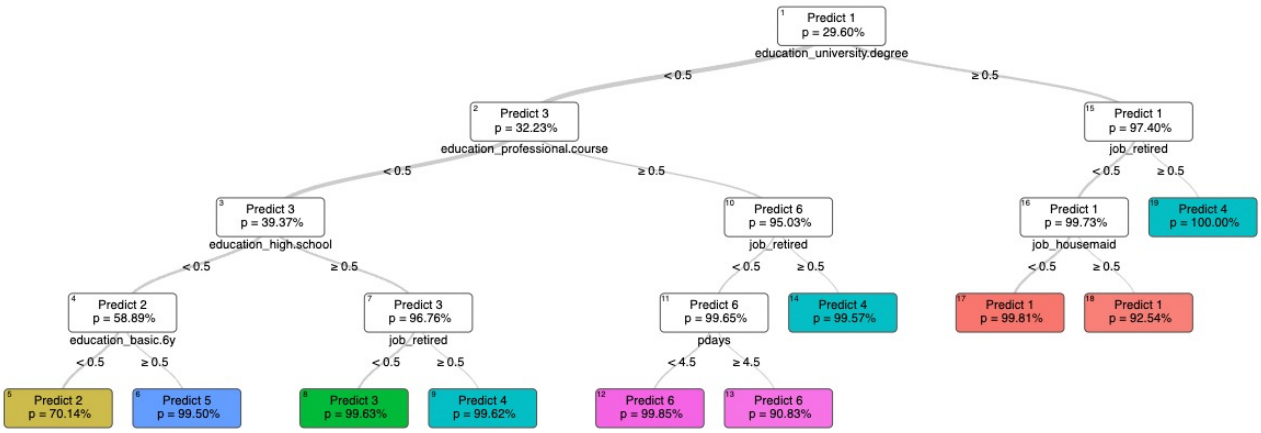
Our dataset, even though relatively clean, still required certain adjustments to assist in model building. Our model did not contain any missing values across rows or columns, but we did identify ~3000 duplicate entries that we decided to drop. In terms of the column values, the "pdays" column which measured the days before the last contact with a user was originally set to 999 in case contact was never made. This value would skew the results for this column and hence we chose to set the value to zero instead which is more reflective of the fact that no contact has been made. Similarly, the outcome variable was transformed to 0 and 1 from "No" and "Yes" respectively.
To deal with the outliers identified during our Exploratory Data Analysis, we applied filtering to our dataset to keep only those customer communications with a duration less than 10 minutes (600 seconds) and those customers that have been contacted less than 5 times.



In terms of numerical variables, we were also cautious of the large difference in scale where certain values were spread across double digits whereas others were in the thousands. We sought to correct for this by standardizing our data by subtracting the mean and then dividing by the standard deviation, which allowed us to simply reduce any bias introduced by differences in scale.
Next, we applied one hot encoding for all the categorical variables of our dataset in order to be able to proceed with predictive methods. The last issue we identified was a large class imbalance in our dataset where the converted user identified by y=1 represents only 9% of the entire dataset and hence we remain wary of this large class difference. Our idea to tackle this, hinges on the use of appropriate performance evaluation metrics whereby we can use metrics such as the AUC score instead of accuracy. For fitting neural networks, we choose to assign class weights in proportion to the number of entries in that particular class to ensure that our model is not biased towards the majority class and the class weights are more equitable distributed to eliminate any inherent bias induced by the data composition.

7 - Interpretable Clustering

Traditional business strategy draws upon a qualitative analysis of the target segment by using business acumen to break down a target market into smaller business segments that can then be captured more effectively. While a qualitative analysis has its merits, through our clustering methods, we underline the value added by a more quantitative approach which is complementary to existing qualitative methods. Our initial approach is to select a subset of the features relevant to the customer and also known to the marketing agent at the time of the call. The first criteria are to simply ensure the clustering incorporates features at the individual level and hence all macroeconomic variables are eliminated due to their inability to offer specific cluster level information. The second layer of filtering ensures that features such as duration are not included, simply owing to the fact that prior to the action of making that call, we cannot reliably estimate call time and hence it is counter intuitive to include it in our clustering feature set.
Our clustering process can be understood across three layers: we initially plot the dissimilarity index from one to ten groups to decide an optimum size for the group keeping in mind the goal of interpretability. Based on this, we identify k=6 as the optimal number of clusters in our model based on the silhouette model. Utilizing this, we then run Optimal Classification Trees with 6 classes to gain an interpretable cluster-level analysis of the variables that define the users allocated to each particular cluster. We utilize the grid search function to allow the OCT to iterate over maximum depth so that the model itself can choose the best set of parameters as opposed to manually setting one. The results of the OCT are as shown below:



In order to demonstrate the power of interpretability exhibited by OCT's, let us break down each cluster in detail:

1. Cluster 1: Individuals with university degrees who are currently active in their jobs
2. Cluster 2: Individuals without any education across all the 4 categories defined for education
3. Cluster 3: Individuals with only a high school education who are still active in their jobs
4. Cluster 4: Three potential routes exist (all for retired individuals):
    a. Retired individuals with only a high school education
    b. Retired individuals with a university degree
    c. Retired individuals without a university degree but with a professional course
5. Cluster 5: Individuals with only a basic 6-year education
6. Cluster 6: Individuals who are currently working without a university degree but with a professional course

For business intuition, clustering users across our target customer base allows us to gain more insight into their general composition and enables the formulation of more targeted outreach strategies. We then implement optimal classification trees on the resulting datasets, and this gives us deep ground- level insight into what factors separate converted users from non-converted ones. Take cluster 6 in the above OCT, we notice that it contains users currently working with professional courses- one user persona here is freelancers so as underlined, this gives us clear business direction.

The OCT above gives us data on an aggregate level but if we desire a more granular level analysis across different classes of users- we can build different OCT's to differentiate users who subscribed to the term deposit from those who did not subscribe to it.

8 - Predictive Methods

This part of our analysis focuses on predicting whether a customer will subscribe to a term deposit or not based on the set of information we possess across the dataset. For our analysis, we employ a wide range of different classification methods that range from traditional methods such as logistic regression and CART to state-of-the-art black box models such as random forest and gradient boosting methods. Regarding evaluation metrics, we remain cognizant of the class imbalance across our dataset and hence lean towards using the area under the curve (AUC) as our metric of choice. Notice that each model we also calculate other metrics such as the F1 score, and the model's accuracy but we chose to compare the models based on AUC. Notice that we run all our models (except OCT-H) for 10 different seeds and report the average performance in order to minimize the effect of randomness.

8.1 - Feature Selection

To begin with, we split our data into training and testing set using stratified sampling, and for each model we utilize a **5-fold cross validation** along with a GridSearch for hyperparameter tuning. Due to the large number of variables that appeared after one-hot-encoding the categorical variables, in order to get rid of noise we utilize feature selection. In particular, we utilize Optimal Feature Selection trees offered by IAI, Using GridSearch we tune our tree to evaluate the sparsity hyperparameter between 1 and 20 nonzero features (out of 57 original). The model selected a sparsity of 13 as the best parameter with the selected features being:

X1 = age	X4 = previous	X7 = cons.conf.idx	X10 = marital_married
X2 = campaign	X5 = emp.var.rate	X8 = euribor3m	X11 = housing_yes
X3 = pdays	X6 = cons.price.idx	X9 = nr.employed	X12 = loan_yes
X13 = poutcome_success			

Below we provide a brief description of the hyperparameters evaluated during the fitting of each distinct model. The evaluation criterion to choose the best hyperparameters is the Area Under the Curve (AUC).

8.2 - Logistic Regression

To tune the logistic regression model, we test its performance without a regularizer but also with the Lasso, Ridge, and Elastic Net regularizers. The best performing model was with the Lasso (l1) regularizer with an out of sample AUC of 83.56%.

8.3 - Decision Trees

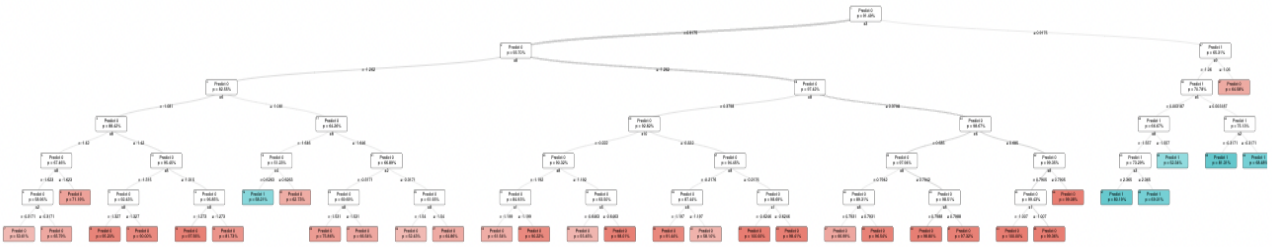
To tune the simple CART model, we test its performance with various hyperparameters using Gridsearch and cross validation. In particular we try various values for maximum depth, minbucket, impurity measures (Gini vs entropy as they handle the class imbalance better compared to the misclassification criterion) and finally the alpha complexity parameter. The best performing tree had depth 5, minbucket 100, entropy impurity and cp value of 0.00024 with an out of sample AUC of 82.31%.

8.4 - Random Forest

To tune the Random Forest model, we test its performance with various hyperparameters by trying various values for maximum depth, number of estimators, and impurity measures (Gini vs entropy). The best performing model had max depth 6, entropy impurity and 100 estimators with an out of sample AUC of 83.39%.

8.5 - Optimal Classification Trees

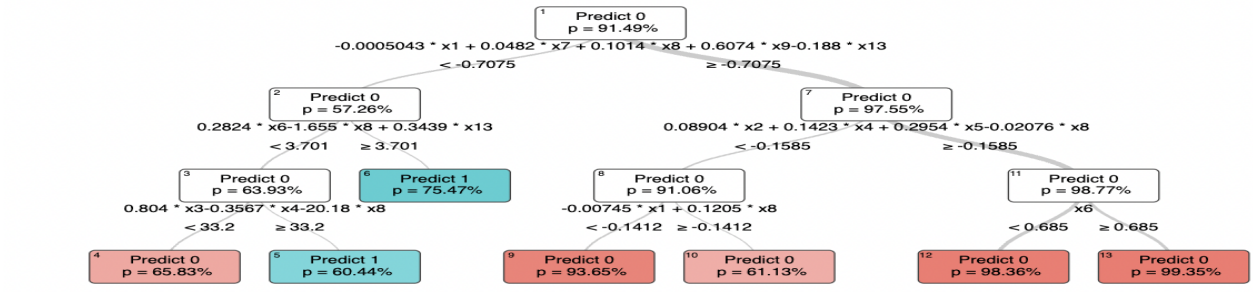
To tune the Optimal Classification Tree model, we utilize the package provided by IAI and we test its performance with various hyperparameters by trying various values for maximum depth, minbucket, alpha complexity parameters and impurity measures (Gini vs entropy). The best performing model had max depth 6, entropy impurity, minbucket 100 and cp value of 0.0006 with an out of sample AUC of 86.71%. A depiction of our tree could be viewed below:



*Note: The OCT's discussed exhibit depth 6 to maximize predictive power which comes at the cost of interpretability- for specific needs, we could reduce the depth.

8.6 - Optimal Classification Trees with Hyperplanes

To tune the Optimal Classification Tree with Hyperplanes model we utilize the package provided by IAI and we test its performance with very few hyperparameters due to its computationally demanding nature.. The best performing model had max depth 3, entropy impurity and cp value of 0.0007 with an out of sample AUC of 86.85%. Notice that we expect the model to perform even better with more tuning, although with the current choice of hyperparameters the model took around 8 hours to train. However, our tree still outperforms all other methods. A depiction of our tree could be viewed below:



8.7 - SVM Model

To tune the Support Vector Machine model, we test its performance with the linear versus the radial basis function kernel. The best performing model was with the RBF kernel with an out of sample AUC of 81.57%.

8.8 - K Nearest Neighbors

To tune the K Nearest Neighbors model we test its performance with various numbers of k neighbors. The best performing model was with the k=500 with an out of sample AUC of 79.24%.

8.9 - Naive Bayes

The simple Naive Bayes model was the worst performing one with an out of sample AUC of 69.69%.

8.10 - Bagging

To tune the Bagging model, we test its performance with various parameters such as the number of estimators, whether to use a warm start or not and whether to use out of bag samples to estimate error. The best performing model was with 1000 estimators, a warm start and out of bag samples with an out of sample AUC of 74.58%.

8.11 - ADA Boost

To tune the ADABoost model we test its performance with various parameters such as the number of estimators and the learning rate. The best performing model was with 1000 estimators and a learning rate of 0.1 with an out of sample AUC of 82.10%.

8.12 - XG Boost

To tune the XGBoost model we test its performance with various parameters such as the number of estimators, the learning rate, the max depth, and the lambda regularization parameter. The best performing model was with 100 estimators, a learning rate of 0.05, the ridge regularizer and a max_depth of 6 with an out of sample AUC of 81.55%.

8.13 - Gradient Boosting

To tune the Gradient Boosting model, we test its performance with various parameters such as the number of estimators, the learning rate, the max depth and the loss parameter. The best performing model was with 100 estimators, a learning rate of 0.05, the log loss and a max_depth of 5 with an out of sample AUC of 81.28%.

9 - Results

The performance of our models could be summarized in the following table:

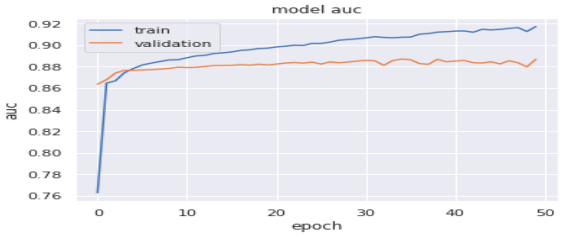
Model	AUC	F1 Score	Accuracy
Logistic Regression	83.56%	94.51%	92.88%
Decision Trees	82.31%	94.36%	92.85%

Random Forest	83.39%	94.48%	**92.92%**
Optimal Classification Trees	86.71%	96.06%	92.59%
Optimal Classification Trees with Hyperplanes	86.85%	**96.14%**	92.62%
SVM	81.575	94.25%	92.85%
K Nearest Neighbors	79.24%	93.95%	92.59%
Naive Bayes	69.69%	89.54%	90.09%
Bagging	74.58%	92.57%	92.07%
ADA Boost	82.10%	94.33%	92.85%
XG Boost	**87.03%**	96.05%	92.60%
Gradient Boosting	81.28%	89.54%	90.09%

From the above results we can immediately notice the Edge of XGBoost and of the Optimal Classification Trees (parallel & with hyperplanes) compared to the other models in terms of AUC. These findings clearly indicate the power of these models. In particular we believe that the Optimal Classification Tree with hyperplanes could improve even more with further tuning, however, it would become computationally unbearable for the purposes of this project. It is worth mentioning that the Logistic Regression model performs extremely well, even when compared to most of the boosted methods.
As a final finding, most of our models identified the feature euribor3m as the most important feature followed by the consumer confidence index and the consumer price index. These results can lead us to believe that the average market situation strongly affects whether a bank customer is willing to make a deposit or not.

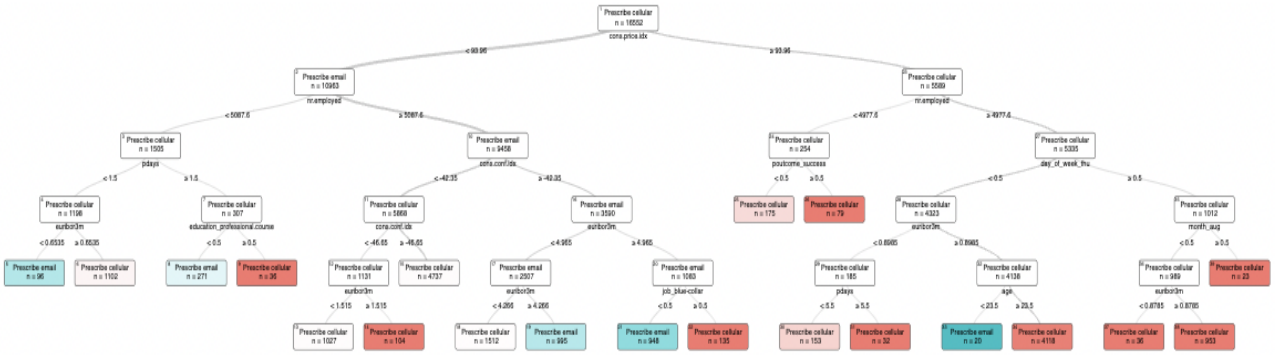
10 - Neural Networks

To further explore our predictive power, we decided to utilize Feed Forward Neural Networks, following the applications we saw in class. Neural networks are state of the art deep learning models that mimic the human brain in terms of decision making by utilizing hidden layers of perceptrons. To tune our neural network, we experiment with various numbers of layers, number of nodes per layer, activation functions in the hidden and the output layer, as well as the learning rate, the batch size and the number of epochs.
After testing various combinations of parameters, we build a neural network with three hidden layers with the ReLu activation function. In terms of class weights, we account for class imbalance by making the weights for each label inversely proportional to the number of entries in each class and in doing so, ensure that equal weightages are assigned to both labels. This makes sure that our model is not biased towards the majority class. We deploy AUC as the metric for our choice in terms of evaluating the performance of our neural network. The best size of epochs was 40 and the best batch size was 1000. Accordingly, the best number of neurons per hidden layer was 256 and the best learning rate was 0.01. For each epoch we utilize a 75/25 training / validation split to better choose our hyperparameters.
In terms of results, our best performing neural network gives an out of sample AUC of 85.21 which outperforms most of the other classification methods we have employed in our analysis but still underperforms when compared to the results of Optimal Classification Trees and XGBoost. This could be due to the need of further tuning in terms of combinations in the number of layers, nodes and activation functions used.



11 - Prescribing the Contact Medium with Optimal Prescriptive Policy Trees

In order to apply the prescriptive methods, we saw in class, we slightly modify our problem. Our ultimate objective here is to decide the best medium to use for contacting a customer in order to maximize the number of people who make a deposit. To achieve that we utilize the feature "contact" which originally had the values "cellular" and "telephone" in a 2:1 ratio, and we just switch the "telephone" value to "email" in order to make our problem more realistic. Our objective now is to maximize the number of deposits by contacting the customer with the right medium. To achieve this objective, we utilize Optimal Prescriptive Policy Trees by IAI, assigning as outcome whether a customer made a deposit, as x the customer features and as treatment the medium used. We split our dataset to a training and testing set in a 50-50 ratio so that we save more data for testing to ensure high-quality reward estimation on the test set, and we set the prescription factor to 0.6 to strike a balance between maximizing the outcome and estimating the approval probability. After tuning using GridSearch the maximum depth and the minbucket parameter we get the following tree:



We immediately notice that our policy tree suggests the best medium to contact mostly based on the consumer price and Euribor indexes, the number of days since last contact, the outcome of the previous campaign offering to the customer, the number of employees working that day, the day of the week, and the occupation and the age of the customer. In prescription problems, it is complicated to evaluate the quality of a prescription policy because our data only contains the outcome for the treatment that was received. To resolve this, we utilize reward estimation, where so-called rewards are estimated for each treatment for each observation. These rewards indicate the relative credit a model should be given for prescribing each treatment to each observation, and thus can be used to evaluate the quality of the prescription policy. Our model's internal outcome estimation models have AUCs of 88% for "cellular" and 82% for "email", which suggests that the reward estimates are of good quality, and good to base our training on. The AUC for the propensity model is at 93%, which is very high, and suggests that our model is reliable in estimating the propensity. The doubly robust estimation method used for fitting the tree helps to tackle such an issue, as it is designed to deliver good results if either propensity scores or outcomes are estimated well.

The average reward achieved through all prescriptions on our test set is 0.12 while the mean reward achieved under the actual treatment assignments was 0.08. We can see that the prescriptive tree policy improves by 50% using our tree model, a significant result on our goal to achieve greater returns from this campaign.

12 - Key Findings

Our key findings can be decomposed across two different sections: the first outlines the promise that interpretable clustering offers in terms of building a cohesive understanding of customer segments and the second underscores the predictive power that modern classification techniques possess and how to harness them in efficient manners. Finally, we showcased how prescriptive approaches could lead to significant improvements on how business is conducted, leading to exceptionally better results.

Interpretable clustering through the use of optimal classification trees can be seen as a powerful tool in breaking down total user base into neatly defined business categories. By closely identifying the variables that determine the composition of each different cluster, we are able to provide actionable insights into how to plan marketing calls for each different segment. Tailoring the product pitch to each different cluster can ensure that marketing agents offer each customer the term deposit in a manner that maximizes their likelihood of conversion and hence leads to higher returns for each dollar spent on marketing. The applications of interpretable clustering can also be seen across different avenues of other industries and companies whereby we can decompose larger customer bases into sub-parts and strategize effectively.

In terms of prediction methods, we can better understand the likelihood of individual customers being converted. Our methods have shown that it is possible to predict this problem to high accuracy levels and hence we remain confident in utilizing such techniques to rethink customer targeting strategies. For example, from a business perspective, we could choose to rank prospective customers according to the likelihood of their conversion and as opposed to randomly calling users, we could focus in descending order on users that based on our model are more likely to convert. This not only means that we are able to convert low hanging fruit in the early phases but has crucial repercussions for cost savings in turbulent economic environments: when banks are looking to curtail marketing expenses, we can offer comparable levels of conversion in terms of absolute number for lesser calls made. This remains conjecture at this point but based on the promise embodied by our models, we are confident that with the right execution, this hypothesis should hold true.

In terms of prescription methods, our focus is to answer another key question for banks, but ultimately for any company that utilizes direct marketing: is there an optimal medium through which we initiate contact with different user types? Through implementing optimal policy trees, we show a 50% improvement in average reward in predictions by optimally deciding whether to call a user or send them an email. This underlines in general business settings the power of understanding what channels of customer acquisition work best for different segments- an issue that issue modern tech businesses hinge on.

13 - Limitations

In terms of evaluating the performance of the models discussed above, we remain confident of the predictive power exhibited by our models but would like to exercise caution in terms of generalizations from the above dataset. Our data sample remains limited to the direct marketing calls undertaken by one specific Portuguese bank, and this restricts our ability to extrapolate findings to other banks. Term deposits are primarily a financial decision in terms of the monetary benefit provided to customers and we did not possess any data to measure the interest rate differentials across different banks. Hence, we cannot safely come to a larger conclusion that could be valid across the banking industry. Our analysis therefore remains limited to the actions of this specific bank, but future studies given access to a wider aperture of data would do well to incorporate parts of our analysis in order to holistically understand the banking term deposit prediction problem in a manner that merges predictive and interpretation power.

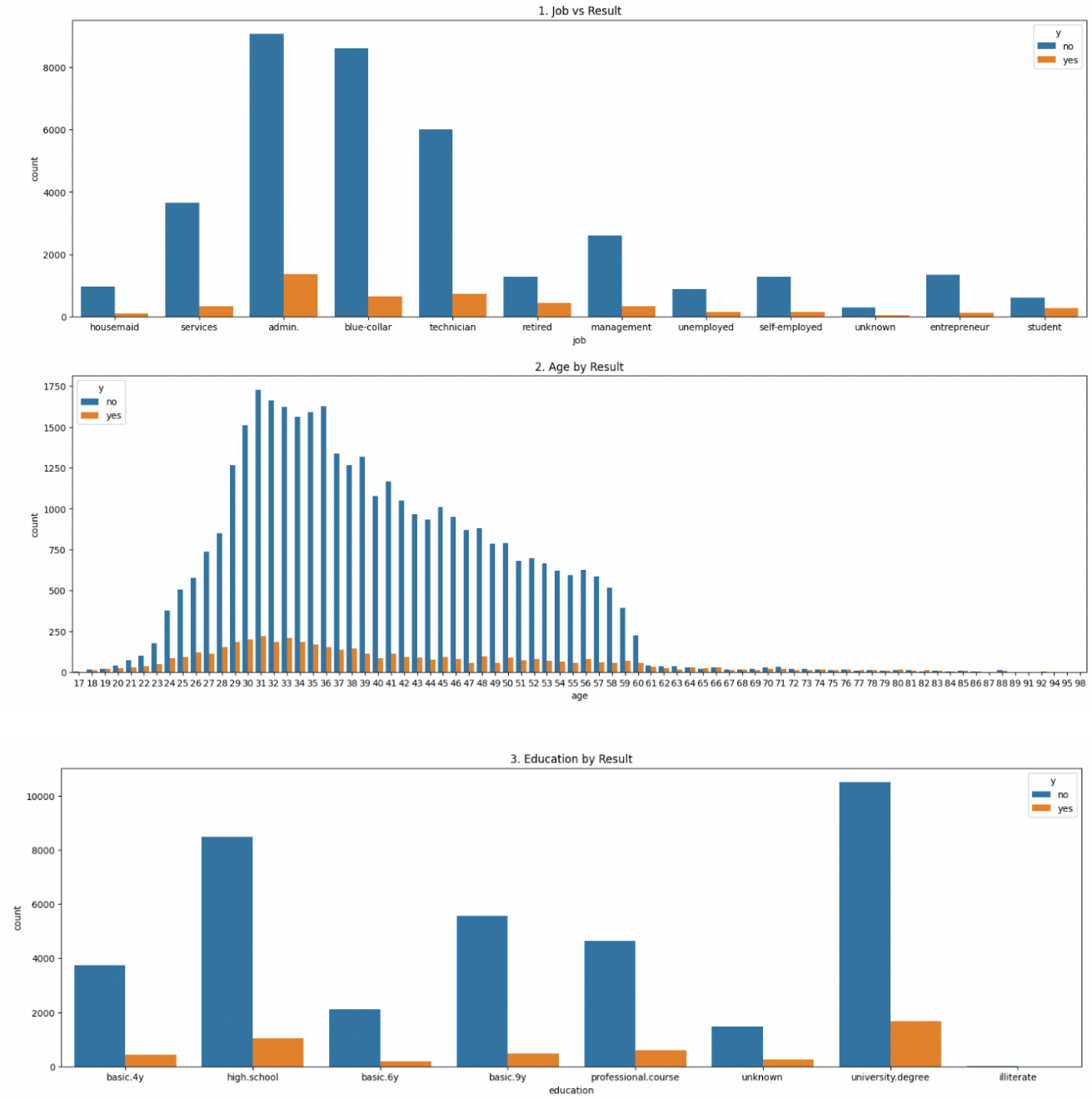
14 – Members Contribution

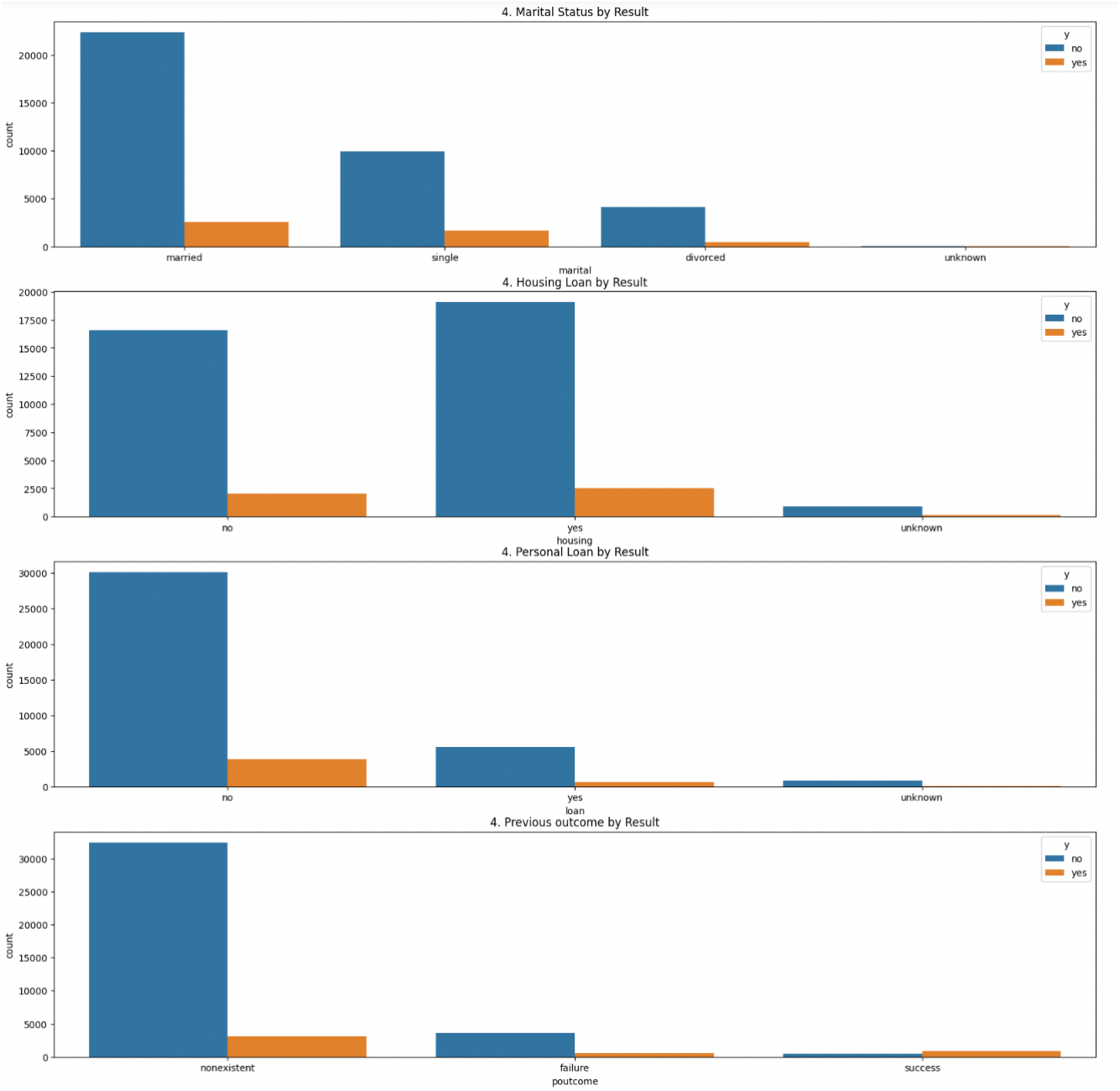
For the purposes of our project, each of our team's members made the following contributions:

- Dataset Identification and Problem Statement → Ahmad, Nikos
- Data Preprocessing → Ahmad
- Exploratory Data Analysis → Nikos
- Interpretable Clustering with OCT → Nikos
- Business Interpretation of resulting clusters → Ahmad
- Feature Selection with OCT → Nikos
- Predictive Methods (fitting and tuning of models) → Nikos
- Neural Network (fitting and tuning) → Ahmad, Nikos
- Prescription Using Optimal Prescription Trees (fitting and tuning) → Nikos
- Final Report Composition → Ahmad, Nikos
- Final Presentation → Ahmad, Nikos

Appendix

In this section we present complementary graphs that further explain the conclusions raised at the Exploratory Data Analysis section regarding the distribution of deposits across people with various characteristics, e.g., age, occupation etc..




4. Marital Status by Result
count
marital
married
single
divorced
unknown
y
no
yes
4. Housing Loan by Result
count
housing
no
yes
unknown
y
no
yes
4. Personal Loan by Result
count
loan
no
yes
unknown
y
no
yes
4. Previous outcome by Result
count
poutcome
nonexistent
failure
success
y
no
yes