Bank Customer Churn Prediction using Decision Trees And Random Forest Ensembles

Pedro Masi

November 8, 2023

1 Introduction

This executive summary provides an overview of our recent consultancy project focusing on determining bank customer churn using a decision tree classification algorithm. The project aimed to develop a decision tree model to assist the bank in making informed decisions based on what type of customers have exited the bank and what type of customer stayed as customers. Preliminary results show that key factors such as age, number of bank products and estimated salary have a high predictive power in determining customer's exits.

2 Objectives

Our primary objectives for this project were as follows:

- To analyze and understand the client's dataset, which contained various categorical and continuous features to ultimate determine the binary target variable which represented customer churn.
- To build an accurate and interpretable decision tree classification model that could predict the target variable based on a robust feature selection procedure and hyperparameter tunning.
- To evaluate the model's performance and identify the specific characteristics of customers that exit the bank.
- To provide actionable recommendations based on the model's outcomes and customer churn characteristics.

3 Findings

Our analysis revealed the following key findings:

- The decision tree model achieved a decent out of sample accuracy of 74 percent, indicating its effectiveness in classifying the bank customer churn on unseen data

- Conducting a feature importance analysis, we found that the most important features when predicting customer churn are customers' Age, number of bank products, estimated salary, and bank balance.
- We found 3 profile of customers that are more prone to exit and leave the bank (summarized in the following section).

3.1 Profile of Customers that Exited

The first group of people that exited are customers that are active members of the bank, they have more than 2.5 financial products with the bank and they are over the age of 41 (N = 211)

The second group of people that exited are customers over the age of 41 whom are not active members of the bank (N = 1287)

Finally, the third group of people that were predicted to exit the bank are customers under the age of 41 but with more than 2.5 financial products with the bank (N = 303).

3.2 Profile of customers that stayed

The customers that stayed as customers are younger customers under the age of 41 and customers that had between 1.5 and 2.5 number of financial products with the bank (N = 2625).

4 Recommendations and Conclusions

Our results show that the two most important features that predict customer churn are Age and the number of financial products that a particular customer has with the bank. Our analysis show that customers over the age of 41 are more prone to exit, and this is more so if in addition, customers hold more than 2.5 financial products with the bank. These two features are predominant whether the customer is an active member of the bank or not. In addition, the customers that are less prone to exit are the younger clients under the age of 41 who hold a healthy number of financial products (between 1 and 2). My general recommendation is to expand the marketing campaign targeting younger working professionals that are able to withstand holding 1 or 2 financial (debt) products with the bank without negatively affecting their repayment potential.

With respect to the relevance and use of the classification algorithm used in this analysis, I recommend the following actions:

- Incorporate the decision tree model into the decision-making process to assist in predicting customer churn.
- Regularly update and retrain the model to adapt to changing circumstances (unseen data).
- Consider deploying the model in a production environment for real-time decision support (if necessary).

5 Annex

This section of the report contains the technical details of the project flow. To support the interpretability of the results, the predictors or features in the data set were not scaled and were left in their original unit values.

5.1 Feature Selection and pre-processing

I used a random forest classifier to help select the most important features in the model. Figure 1 shows the ranked feature importances for customer churn. Prior to fitting the model, I realized that the target class (binary exited) was highly unbalanced where 7963 observations were non-exited instances and 2037 observations were exited instances. To prevent the decision tree to be biased towards the dominant class (non-exited), I oversampled the dataset so both classes were equally represented at the root or beginning of the tree.

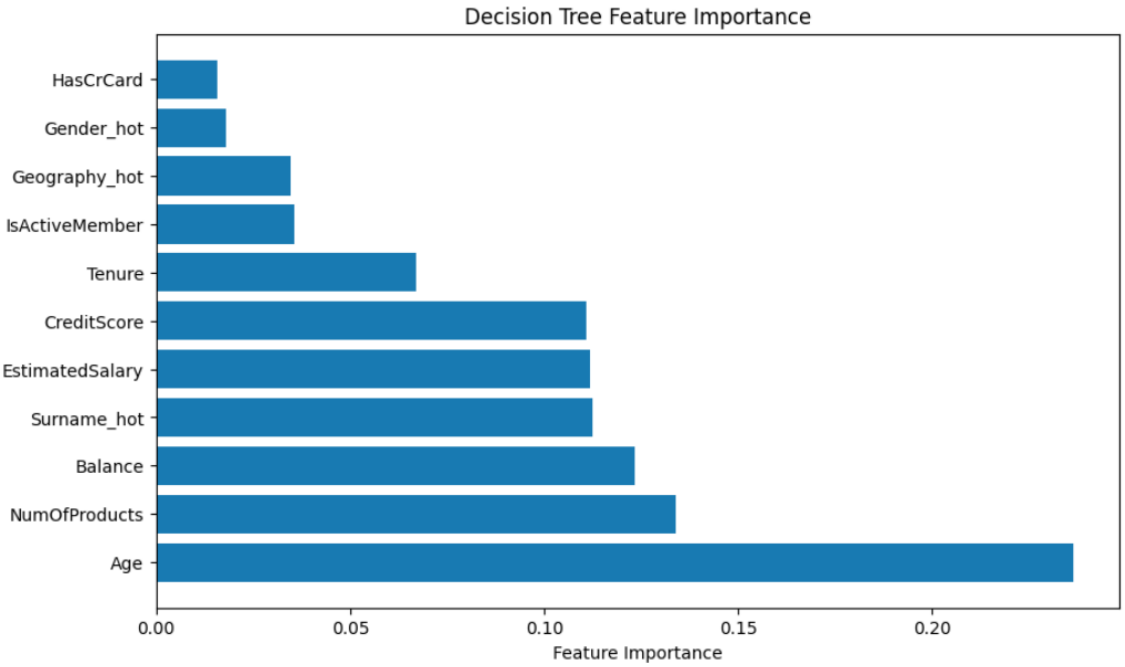

Figure 1: Ranked Feature Importance for Bank Customer Churn

5.2 Hyperparameter considerations

Prior to training the decision tree model, I conducted a grid-search to determine the optimal value for the depth and number of leafs in the tree. The results of the grid-search showed the maximum depth of the tree of 6 and a maximum number of leaf nodes of 8. Given the fact that the out of sample accuracy score of a decision tree with depth of 6 and depth of 4 was almost identical

(approximately around 73.3 percent), I decided to use a lower depth to keep the tree simple and interpretable. Figure 2 shows the decision tree of the model.

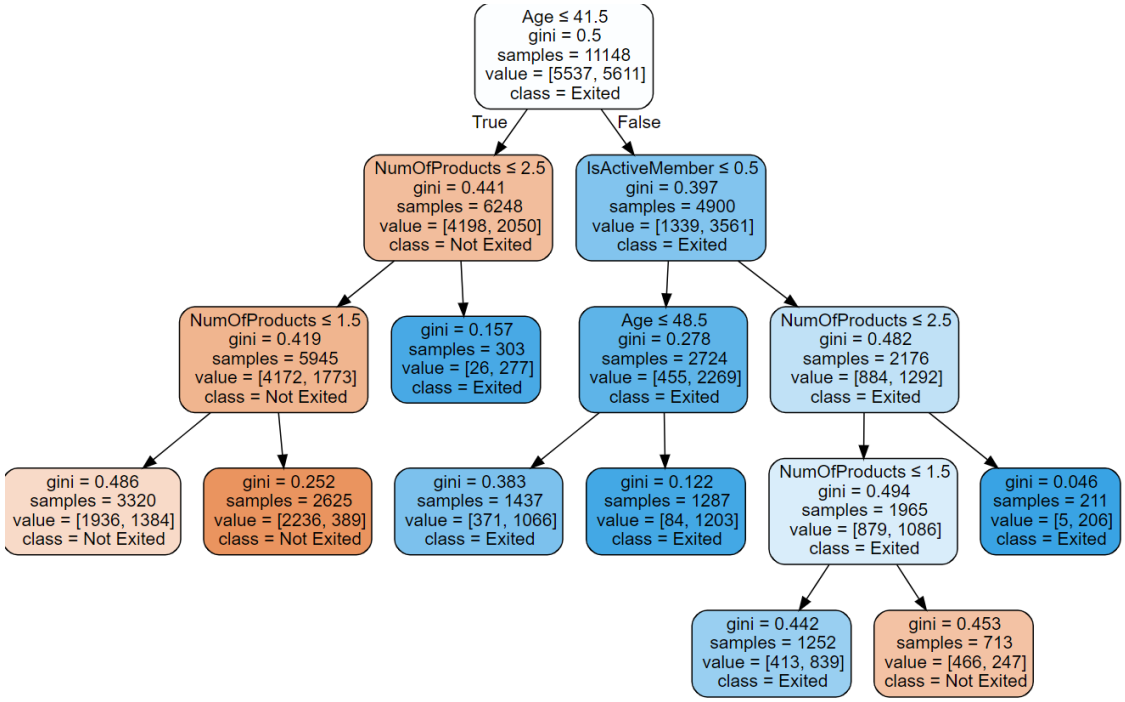

Figure 2: Decision Tree Graph

5.3 Additional considerations and Ensemble Model

The use of the decision tree model helped make interpret-able and insightful recommendations for executives at the bank related to customer churn. However, the out of sample performance of the decision tree could have been improved. To improve the accuracy of the model, I decided to fit a ensemble model (random forest) to see if my predictions were better. I performed a grid-search and determined the optimal hyper-parameter was to have a max depth of 50 while leaving all other hyperparameters unbounded. As a result, I got a significant improvement in my out of sample performance (94.3 percent). Nevertheless, this comes at a cost, as seen by Figure 3, we are no longer able to clearly interpret the ensemble model. As a result, we are able to imporove our predictions but we cannot clearly explain the rules by which customers churn.

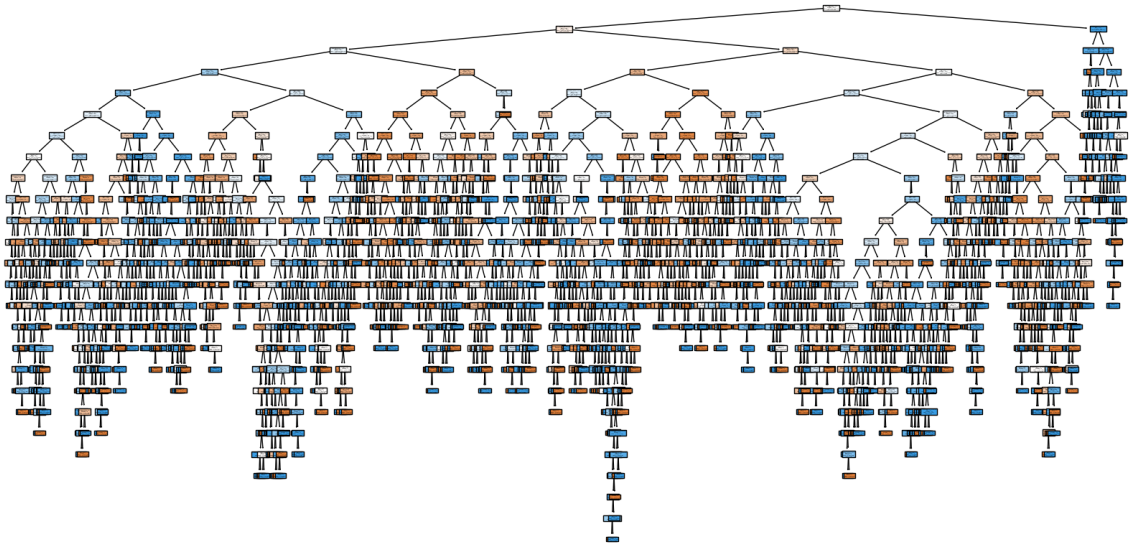

Figure 3: Ensemble Random Forest Tree 1/50