Banking Customer Churn - Prediction

Milestone 4: Project-2 Report

Vinay Nagaraj

DSC680, Summer 2021

Bellevue University, NE

https://vinaynagaraj88.github.io/DataScience_Portfolio/

Abstract

Customers are the most important part of your business regardless of the industry. There would be no sales without customers and they are a critical factor when developing your marketing messaging and strategy. Customer Churn is the rate at which customers stop doing business with an entity. This is one of the most acknowledged problems in the banking sector and Banks are constantly looking at data/suggestions which could help them improve their customer service and retain their existing customers and also bring in new customers.

Introduction

A customer's banking relationship includes key journeys that range from onboarding and transacting to maintenance and problem resolution. Customers are central to a wave of new opportunities and challenges facing banking executives, with regulators increasingly expecting banks to deliver on more than just credit-risk management and associated capital requirements.

Banks often use customer churn analysis and customer churn rates as one of their key business metrics because the cost of retaining existing customers is far less than acquiring a new one. Customer churn prevention is one of the deciding factors when it comes to maximizing the revenues of any organization.

Through this project, I intend to focus on the behavior of bank customers who are more likely to leave the bank. I want to find out some striking behaviors of customers

through Exploratory Data Analysis and later on use some of the predictive analytics techniques to determine the customers who are most likely to churn.

Methods

The project will be carried out by utilizing the CRISP-DM model. It stands for Cross Industry Standard Process for Data Mining. The process contains 6 steps that will be followed throughout the project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation
6. Deployment

As part of this paper, first 4 steps (mentioned above) will be covered in this methods section. Evaluation and deployment steps are detailed in the results and conclusion section.

- Business Understanding - According to a study by management consultancy cg42, the 10 largest retail banks in the U.S. are in trouble. Based on a survey encompassing more than 3,000 of their current customers, cg42 ranked banks based on projected customer attrition and potential revenue loss. The study, which is an update of research cg42 repeated in 2011 and 2013, found that up to 23% of current bank customers are ready to change banking providers. cg42

says 8% will actually follow through and make the switch [2]. The banks today aren't solving the problem of churn effectively. They continue to lose premium customers, denting revenue and profitability severely. This is clearly a major issue. Big Banks Risk Losing Billions from Disgruntled Customers. As part as of this project, we are trying to identify the factors that lead to customers switching banks and build a classifier model that would help banks in predicting the customer who are very likely to quit the bank.

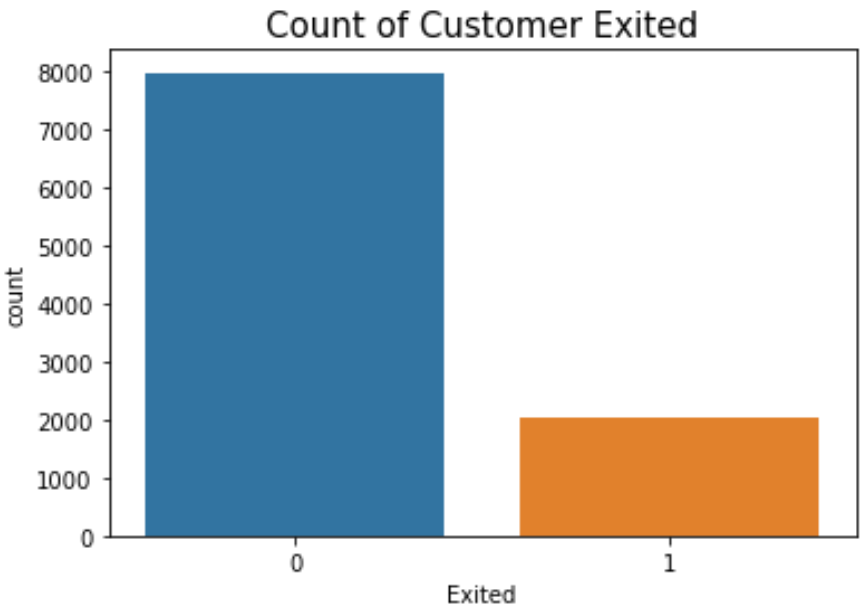
- Data Understanding - I have used the below dataset from Kaggle which contains details of a bank's customers and the target variable is a binary variable reflecting the fact whether the customer left the bank (closed his account) or he continues to be a customer. I am planning to use this data to train the model as part of this project.

  Dataset Link - https://www.kaggle.com/shrutimechlearn/churn-modelling

  The dataset contains 10,000 records with 13 attributes and one target variable. Demographically the information is about customers from Spain, France, and Germany. 55% of customers are Male, and 45% are Female, with an average age of 38.9 years. Column "Exited" looks like the feature I will have to predict which will tell us whether the customer closed his account or not. So, feature “Exited” will mostly be my target variable.

  As part of this step, we performed graph analysis to see the distribution of the variables.

  Fig 1: Below plot shows the count of Customers in our dataset.



Total count of Customers who have an active account is 7,963

Total count of Customers who no longer are active with the bank is 2,037

Fig 2: Below plots shows the gender wise customer churn. Female customers tend to exit the bank more than Male customers. Banks need to see if they can introduce a product or two which can be very beneficial and enticing to the female customers.

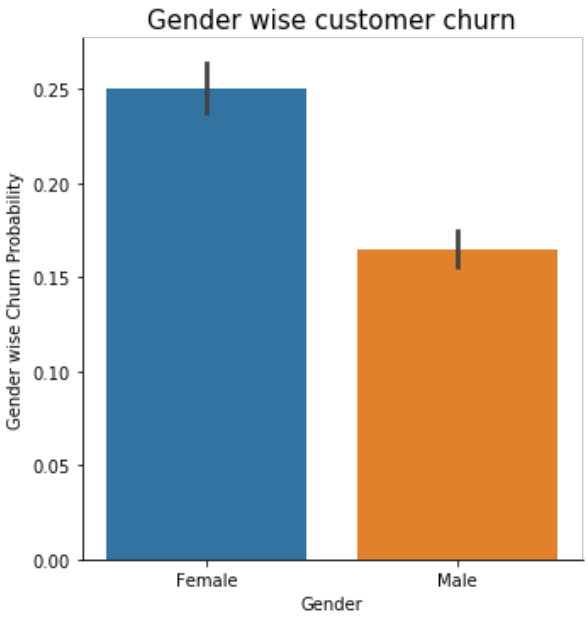


Fig 3: Below plots shows the geography wise customer churn. Customers from Germany tend to exit the bank more than Spain or France. Bank need to improve their customer experience in Germany. They need to see what are the things they are doing better in Spain and France when compared to Germany.

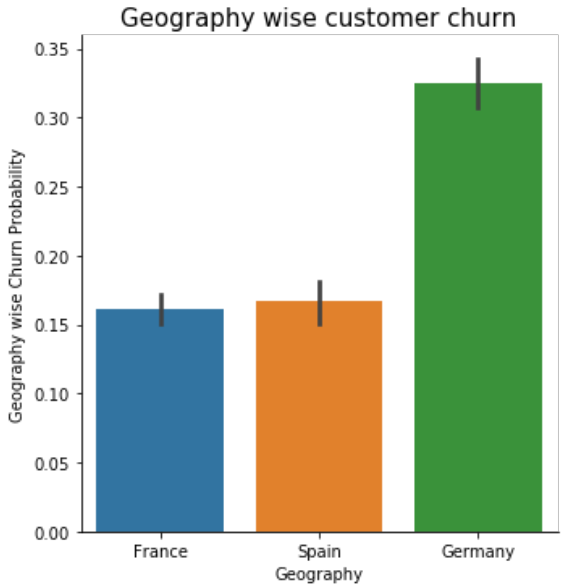


Fig 4: Below plots shows customer churn based on how active the customers are. Inactive customers tend to exit the bank more. It is important to keep the customers engaged. Direct deposits can be a good example.

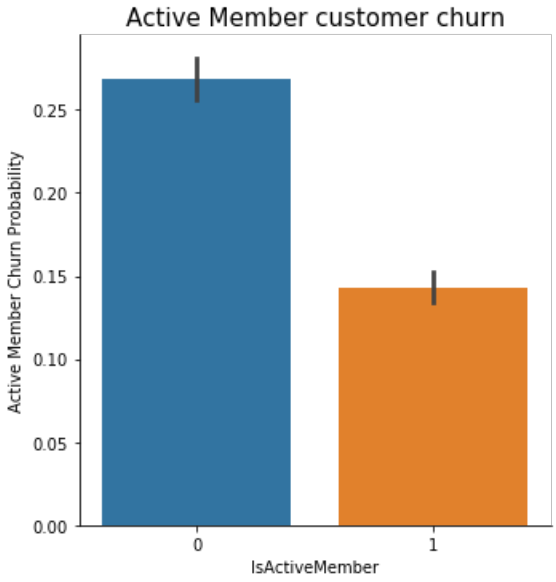


Fig 5: Below plots shows customer churn based on whether they have a credit card or not. Customers with no credit card tend to exit the bank more. Provide welcome bonus and few cash back options to have customers open credit card with your bank.

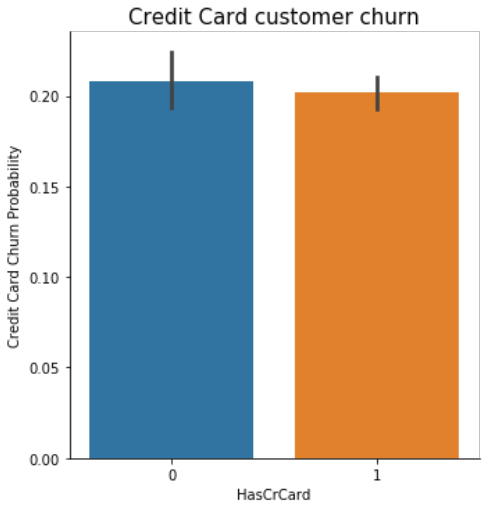


Fig 6: Below plots shows product wise customer churn. The ratio of exited cases with 3 or more products definitely higher than under 2 products. Banks need to review how their products work together.

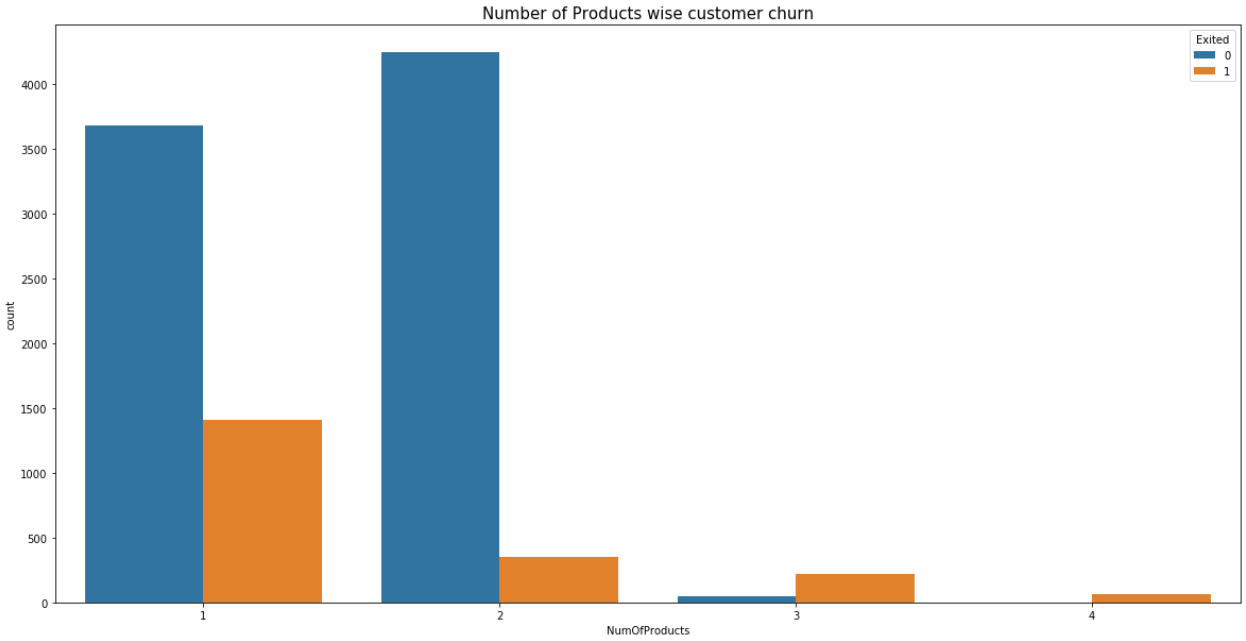
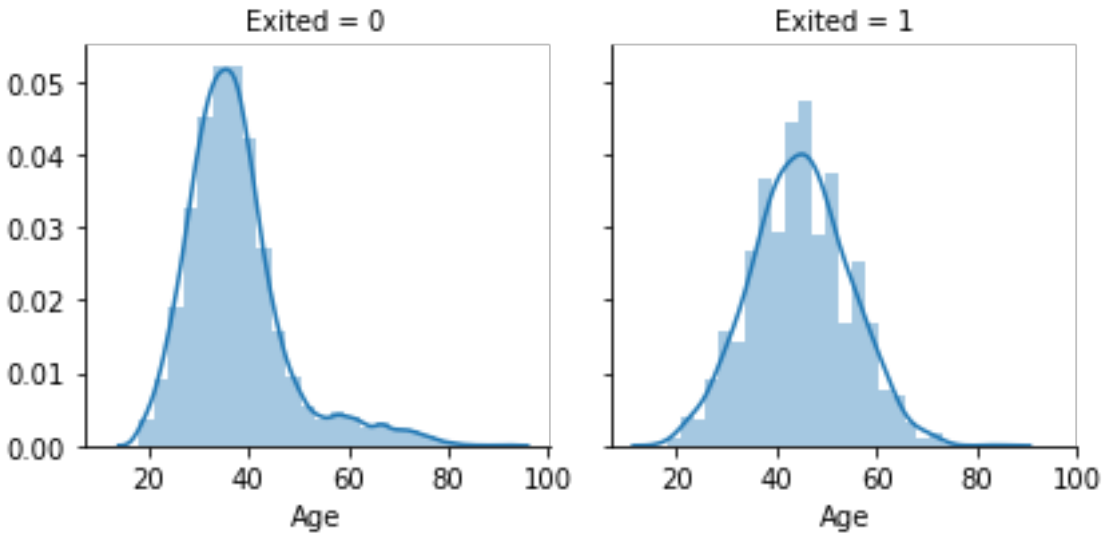


Fig 7: Below plots shows age wise customer. People with Ages between 30 to 40 has the highest probability of staying and Ages between 45 to 55 has the highest probability of leaving. Banks need to review their retirement benefits and other options which interest their aging customers.



- Data Preparation – The dataset I downloaded from Kaggle was a clean dataset with no missing values or outliers. As part of the data preparation step, I understood the characteristics of each feature in my dataset and also in preparation for modeling, I modified all the categorical features to contain numeric values as they were also important features which increased the accuracy of the classification models.
- Modeling – As the first step towards modeling, I used the SelectKBest technique to understand the importance of the features in our dataset. The SelectKBest scores of all the features of our dataset can be found in Appendix A and the plot that shows the features in the order of their importance can be found in Appendix B.

  Using the top 10 features determined by SelectKBest technique, we normalized the data and performed model comparison.

	roc_auc
RandomForestClassifier	0.847526
DecisionTreeClassifier	0.688952
SGDClassifier	0.664615
LogisticRegression	0.752895
KNeighborsClassifier	0.781499
SVC	0.822679
GaussianNB	0.803606

Based on the roc_auc values for all the 7 models we compared, RandomForestClassifier, SVC(Support Vector Classifier) and GaussianNB (Gaussian Naive Bayes) had the better scores and I decided to pursue these models further.

Models – HyperParameter Tuning

Based on the above roc_auc scores, I decided to perform HyperParameter Tuning on Random Forest, Support Vector Classifier and Gaussian Naive Bayes models.

1. **Random forest model** creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object. It provides us a high true prediction values for our dataset. We tested this model on the test dataset.

This model has an AUROC of 0.861 which means that the model has very good discriminatory ability. 2,334 cases are true negative, meaning they are active customers and model correctly predicted them as active customers. 268 cases are true positive, meaning they were predicted as customers who exited the bank and are truly the customers who exited. 316 instances are false negative and 72 are false positives.

```
Confusion Matrix - Random Forest
[[2344   72]
 [ 316  268]]

Classification report - Random Forest
              precision    recall  f1-score   support

           0     0.8812    0.9702    0.9236      2416
           1     0.7882    0.4589    0.5801       584

    accuracy                         0.8707      3000
   macro avg     0.8347    0.7146    0.7518      3000
weighted avg     0.8631    0.8707    0.8567      3000

Scalar Metrics - Random Forest
        AUROC = 0.8610
```

2. **Support Vector Classifier model** is to fit to the data you provide, returning a best fit hyperplane that divides, or categorizes, your data. From there, after getting the hyperplane, you can then feed some features to your classifier to see what the predicted class is. This model has an AUROC of 0.8216 which means that the model has very good discriminatory ability. 2,372 cases are true negative, meaning they are active customers and model correctly predicted them as active customers. 217 cases are true positive, meaning they were predicted

as customers who exited the bank and are truly the customers who exited. 367 instances are false negative and 44 are false positives.

```
Confusion Matrix
[[2372   44]
 [ 367  217]]

Classification report
              precision    recall  f1-score   support

           0     0.8660    0.9818    0.9203      2416
           1     0.8314    0.3716    0.5136       584

    accuracy                         0.8630      3000
   macro avg     0.8487    0.6767    0.7169      3000
weighted avg     0.8593    0.8630    0.8411      3000

Scalar Metrics
        AUROC = 0.8216
```

3. **Gaussian Naive Bayes model** is a variant of Naive Bayes that follows Gaussian normal distribution and supports continuous data. This model has an AUROC of 0.8085 which means that the model has very good discriminatory ability. 2,356 cases are true negative, meaning they are active customers and model correctly predicted them as active customers. 147 cases are true positive, meaning they were predicted as customers who exited the bank and are truly the customers who exited. 437 instances are false negative and 60 are false positives.

```
Confusion Matrix - GaussianNB
[[2356   60]
 [ 437  147]]

Classification report - GaussianNB
              precision    recall  f1-score   support

           0     0.8435    0.9752    0.9046      2416
           1     0.7101    0.2517    0.3717       584

    accuracy                         0.8343      3000
   macro avg     0.7768    0.6134    0.6381      3000
weighted avg     0.8176    0.8343    0.8008      3000

Scalar Metrics - GaussianNB
        AUROC = 0.8085
```

Results

While banks are always on a lookout for new customers, it is very important to make sure to keep the existing customer base happy. Customer retention increases banks customers' lifetime value and boosts banks revenue. It also helps the bank to build amazing relationships with the customers. They trust the bank with their money based on the value they receive in exchange. Existing customers can drive repeat business, increase revenue, create brand ambassadors, defend against competition and gain valuable feedback.

- Banks need to see how they can engage the older generation more with their bank.
- It is important to make sure banks keep the customers actively engaged with the bank.

- Bank need to improve their service to Female customers and also customers in Germany.
- Provide interesting perks to customers who open a credit card with their bank.

Acknowledgements

I'm indebted to the communities behind the multiple open-source software packages on which we depend. I would like to thank my family for their understanding of our time in this endeavor. Last by not least, thanks to Prof Catie Williams and my classmates for their guidance and feedback.

References

[1] Shruti_Iyyer. (2019, April 3). *Churn Modelling*. Kaggle. https://www.kaggle.com/shrutimechlearn/churn-modelling

[2] Jeffry Pilcher, C. E. O. P. and F. of T. F. B. (2020, August 5). Big Banks Risk Losing Billions From Disgruntled Customers. The Financial Brand - Banking Trends, Analysis & Insights. https://thefinancialbrand.com/55748/banking-consumers-attrition-loyalty-study/

[3] Why Customers Leave & What Can Banks Do? Tiger Analytics. (2020, September 16). https://www.tigeranalytics.com/blog/addressing-customer-churn-in-banking/

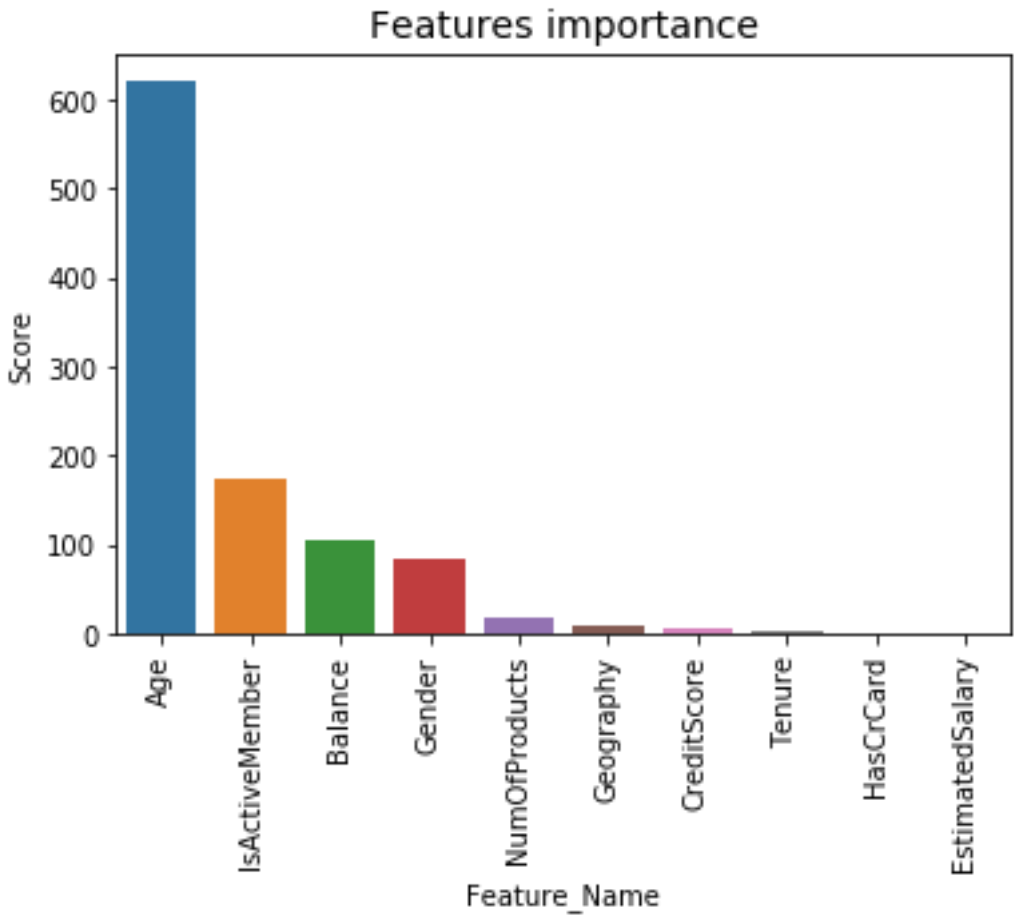
[4] Mukhtar, N. (2020, January 30). Creating a Banking Customer Churn Model. Medium. https://medium.com/@noah.fintech/creating-a-banking-customer-churn-model-1a2d0850f071

Appendix

Appendix A:

	Feature_Name	Score
3	Age	621.256581
8	IsActiveMember	174.139240
5	Balance	104.241621
2	Gender	85.618179
6	NumOfProducts	19.021790
1	Geography	9.799450
0	CreditScore	5.859852
4	Tenure	2.764884
7	HasCrCard	1.878124
9	EstimatedSalary	0.247795

Appendix B:


Features importance
Score
600
500
400
300
200
100
0
Age
IsActiveMember
Balance
Gender
NumOfProducts
Geography
CreditScore
Tenure
HasCrCard
EstimatedSalary
Feature_Name