Customer Churn Prediction in Banking (CCPB)

Aditya Sumbaraju

Bellevue University

DSC680- Applied Data Science

Dr. Brett Werner

01/23/2022

https://github.com/adityasumbaraju/aditya_portfolio/tree/main/Customer_Churn_Prediction_in_banking

Business Problem

Customer churn exists across businesses in many sectors, especially since it significantly impacted banking. A company's growth depends on the high acquisition and low attrition rate. A high attrition rate represents a considerable investment loss, and both time and effort need to be channeled into replacing them. Predicting when a client is likely to leave and offering them incentives to stay can offer considerable savings to a business.

Predicting churn (attrition) is essential for any current subscription-based business. A slight fluctuation in churn can significantly impact the bottom line of any business. Hence it is vital to know- "Is this customer going to leave us within X months?" Yes or No?

Background

Customer Churn, in other terms, is also known as customer turnover, customer defection in the loss of clients or customers, or customer attrition.

Banking is the only service company among several other service-based companies that often use customer churn analysis as one of their key business metrics because retaining existing customers is far less than acquiring a new one.

When new customers subscribe to the scheme in a bank, each new subscriber contributes to the bank product's growth rate. In a nutshell, some of those customers may stop utilizing the scheme and end their subscription. It may be because they could have switched to a competitor and the customer is no longer in need of the bank services. The potential reasons are they're unhappy with their banking experience, or it could be too expensive to afford. The scheme's

subscribers that stop using the bank products are categorized as "churn" for a given period. Churn can be evaluated monthly, quarterly, or annually.

EDA focuses on the behavior of bank customers who are more likely to exit the bank. I want to discover customer behaviors through Exploratory Data Analysis and use machine learning techniques to evaluate the customers who are most likely to churn.

Data Explanation

The data used in this use case to perform EDA and predictive Modeling of customer churn in banking was sourced from Kaggle.

Data Source Urls:

https://www.kaggle.com/mathchi/churn-for-bank-customers

The dataset consists of 10000 observations and 12 variables.

- Independent variables contain information about customers.
- The dependent variable refers to customer abandonment status.

Variables:

- **RowNumber**: corresponds to the record (row) number and does not affect the output.
- **CustomerId**: Contains customer_ids of the bank.
- **Surname**: Surname is considered PII of the customer. Data needs to be Masked for ethical considerations or removed if there is no significance.
- **CreditScore**: This variable will significantly affect customer churn; the higher the credit score, the fewer chances to exit the bank.
- **Geography**: Location of Customer can be a potential churn factor.

- **Gender**: An interesting variable to explore and identify the churn factor.
- **Age**: Older customers are less likely to leave their bank than younger ones, and it contains the customer's Age.
- **Tenure**: This variable signifies customer loyalty. Ideally, loyal customers are less likely to leave, and loyalty is gauged based on the term of stay.
- **Balance**: This variable provides a savings account balance
- **NumOfProducts**: Number of products that a customer has purchased through the bank during their tenure.
- **HasCrCard** : Boolean variable(0=No,1=Yes). It signifies whether or not a customer has a credit card. The hypothesis is people with a credit card are less likely to leave the bank.
- **IsActiveMember** : Boolean variable(0=No,1=Yes). Inactive customers are more likely to leave the bank. This is an important variable for our prediction use case.
- **EstimatedSalary**: Customers with lower salaries/balances are more likely to leave the bank than those with higher wages.
- **Exited**: Boolean variable(0=No,1=Yes) Does the customer leave the bank?

```
#information about the data
data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   RowNumber        10000 non-null  int64
 1   CustomerId       10000 non-null  int64
 2   Surname          10000 non-null  object
 3   CreditScore      10000 non-null  int64
 4   Geography        10000 non-null  object
 5   Gender           10000 non-null  object
 6   Age              10000 non-null  int64
 7   Tenure           10000 non-null  int64
 8   Balance          10000 non-null  float64
 9   NumOfProducts    10000 non-null  int64
 10  HasCrCard        10000 non-null  int64
 11  IsActiveMember   10000 non-null  int64
 12  EstimatedSalary  10000 non-null  float64
 13  Exited           10000 non-null  int64
dtypes: float64(2), int64(9), object(3)
memory usage: 1.1+ MB
```

Fig 1: Metadata

Implementation Plan

The implementation plan contains the below steps to accomplish the project

- Business understanding
- Data Preparation
- Exploratory data analysis
- Data prep for Modeling
- Modeling
- Model evaluation

Methods

I have implemented "standardization" on some features using the sklearn library (StandardScalar). I have scaled down features into properties of Standard Normal Distribution where standard deviation = 1 and mean = 0. The takeaway observation is scaling gave a higher performance in classifiers such as Random Forest, Logistic Regression, KNN, Support Vector Classifier. It proves that feature scaling improved the performance of mentioned classifiers to predict an accurate model using the CRISP-DM method. Below are the phases I have targeted to achieve a better model.

- **Business understanding**: I would be focusing on case study objectives and requirements from a business perspective. This is a valuable phase to create a preliminary plan and move with the subsequent stages.
- **Data Understanding**: Dataset is already identified to proceed with the case study. All I need to do is identify data anomalies detect interesting subsets from the hypothesis.

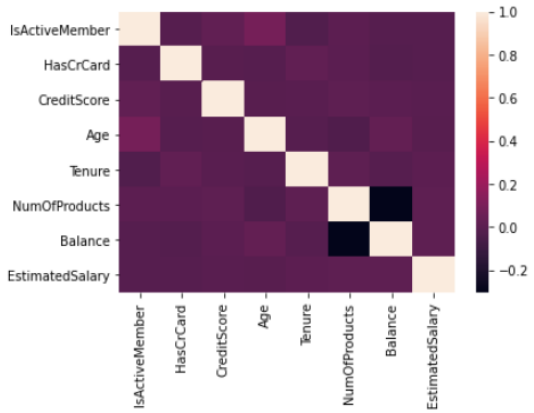
- **Data Preparation**: This phase covers the data wrangling scenarios to construct the final dataset from the raw dataset.
- **Modeling and evaluation** involve building and developing various models based on different modeling techniques. We determine the predictive modeling algorithm in this stage and evaluate which models give the best performance. Pending modeling selection and approach, we might need to split the data into training and test sets using the sklearn train_test_split library. Since the project is classification-based (exited or not exited), I used classification models such as Logistic Regression, Support Vector Classifier, KNN, Randon Forest to make a prediction. After evaluating several operations such as hyper-parameter tuning cross-validation, the highest accuracy was recorded with Random Forest Classifier followed by SVM. Check the Model Evaluation section to see how each classifier performed.

- **Deployment**: Publish final results and conduct a case study retrospect on what went well and required to build the Model better.

Exploratory Data Analysis

```
*********Analyze Correlation between numerical feature using Heatmap plot****************
```

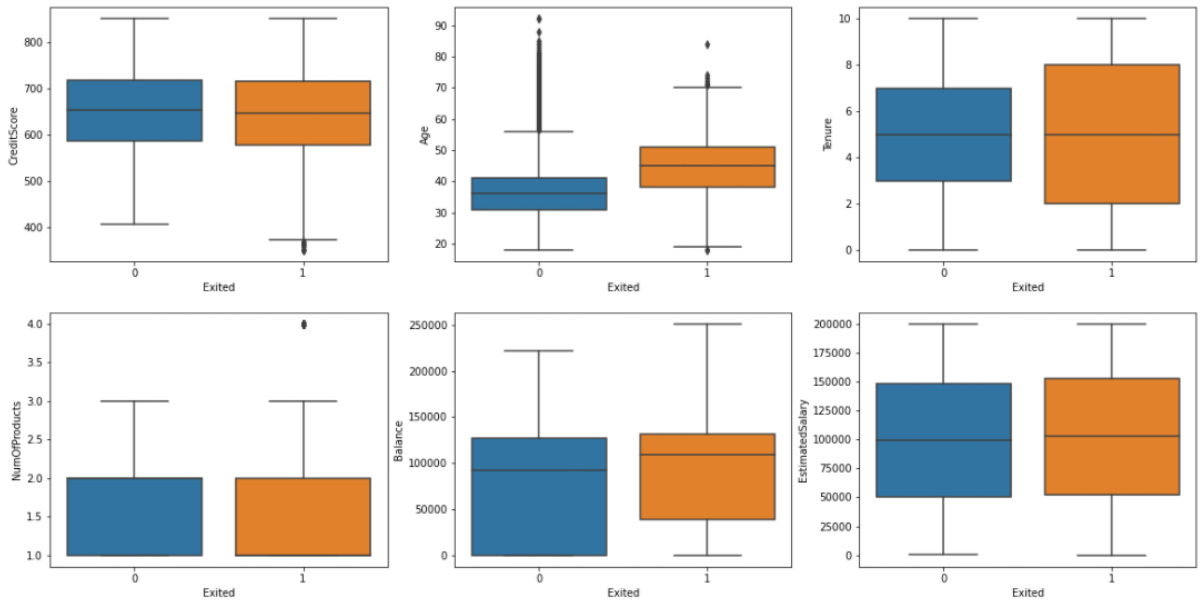
```
<AxesSubplot:>
```



Correlation Matrix

```
*********boxplot for numerical features****************
```

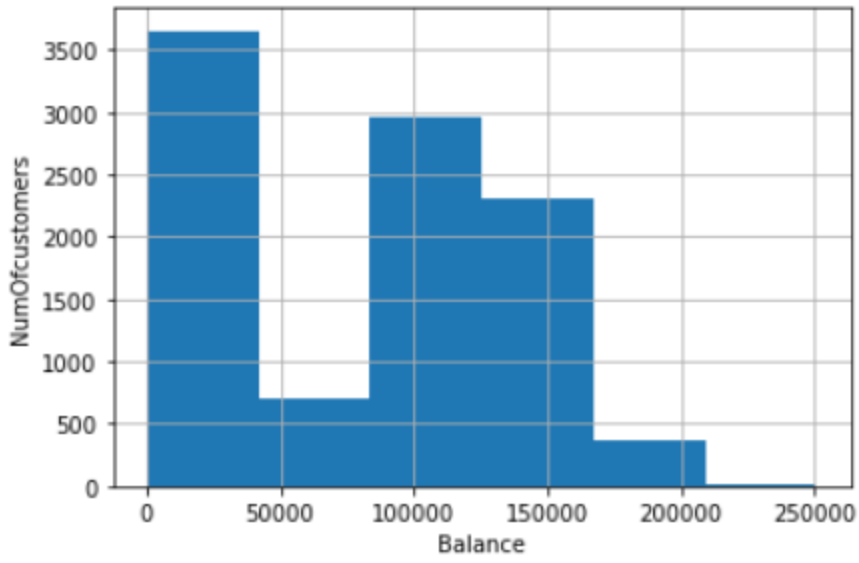
```
: <AxesSubplot:xlabel='Exited', ylabel='EstimatedSalary'>
```



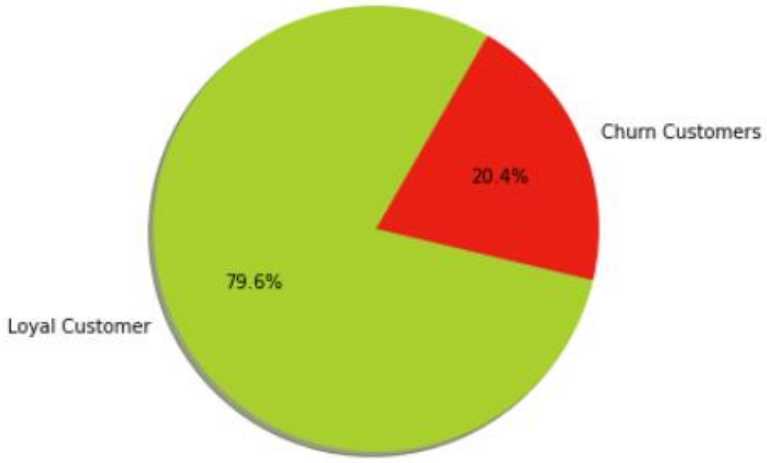
Numerical Features

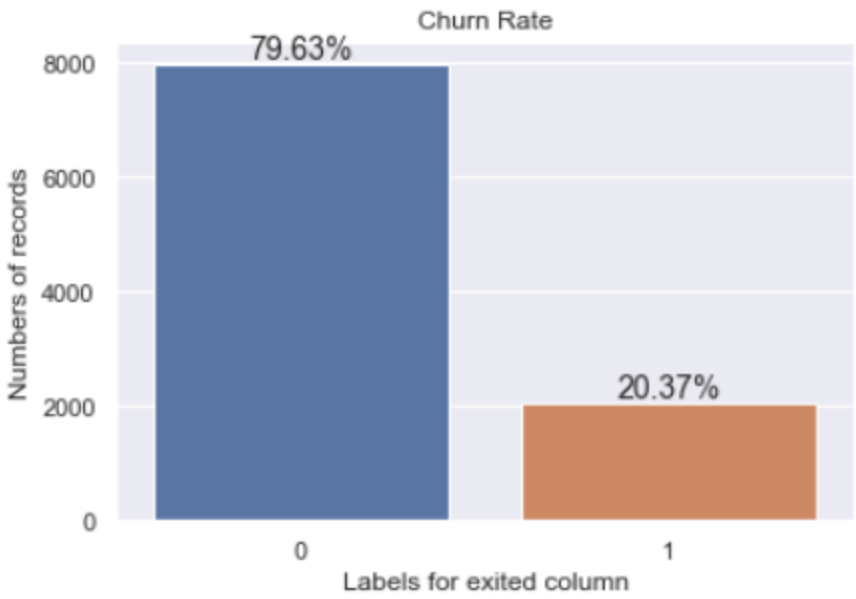
**********Balance Distribution EDA**********

```
Text(0, 0.5, 'NumOfcustomers')
```



**********What is the Percentage of loyal customers vs churn customers?**********





ChurnRate

```
ccpb_df_cleaned.groupby(ccpb_df_cleaned["Exited"])["Age"].mean()
```

```
Exited
0    36.089197
1    43.793583
Name: Age, dtype: float64
```

Chrun distribution w.r.t Age

Modeling Results

Logistic Regression:

	precision	recall	f1-score	support
0	0.85	0.96	0.90	2294
1	0.64	0.30	0.41	561
accuracy			0.83	2855
macro avg	0.74	0.63	0.66	2855
weighted avg	0.81	0.83	0.80	2855

Model Accuracy score: Logistic Regression 0.8294220665499125

SVC:

	precision	recall	f1-score	support
0	0.86	0.97	0.91	2294
1	0.77	0.36	0.49	561
accuracy			0.85	2855
macro avg	0.82	0.67	0.70	2855
weighted avg	0.84	0.85	0.83	2855

Model Accuracy score: Support Vector Classification 0.8539404553415061

Random Forest:

	precision	recall	f1-score	support
0	0.87	0.96	0.92	2294
1	0.74	0.42	0.54	561
accuracy			0.86	2855
macro avg	0.81	0.69	0.73	2855
weighted avg	0.85	0.86	0.84	2855

Model Accuracy score: Random Forest Classifier: 0.8574430823117338

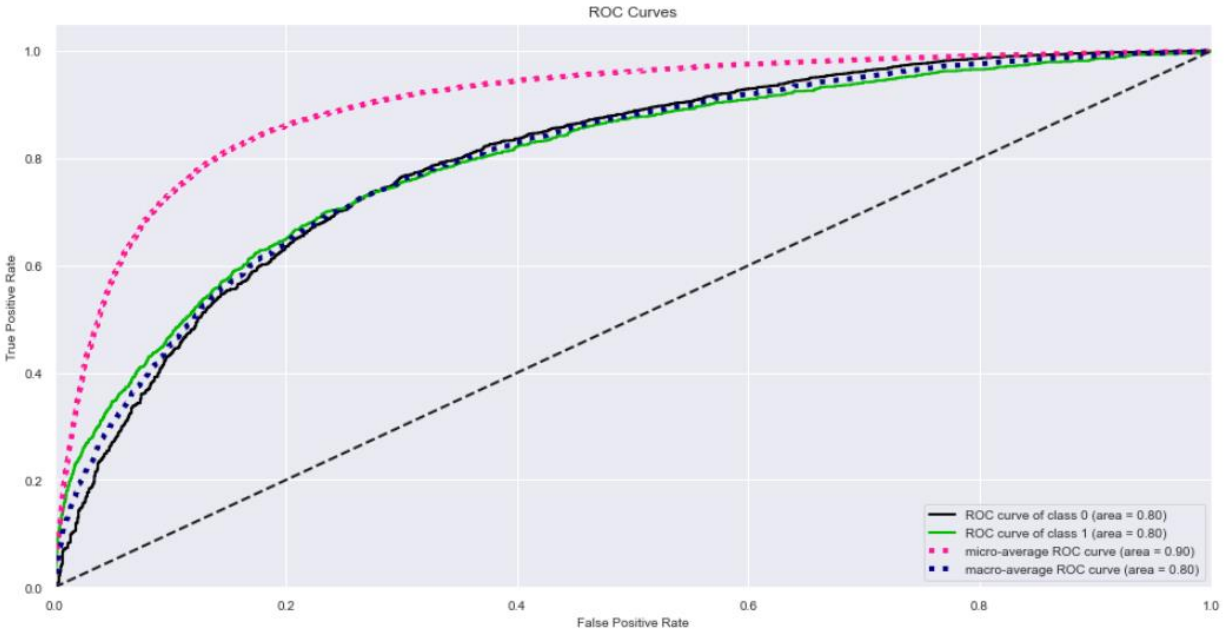
KNN:

	precision	recall	f1-score	support
0	0.85	0.95	0.90	2294
1	0.61	0.34	0.44	561
accuracy			0.83	2855
macro avg	0.73	0.64	0.67	2855
weighted avg	0.81	0.83	0.81	2855

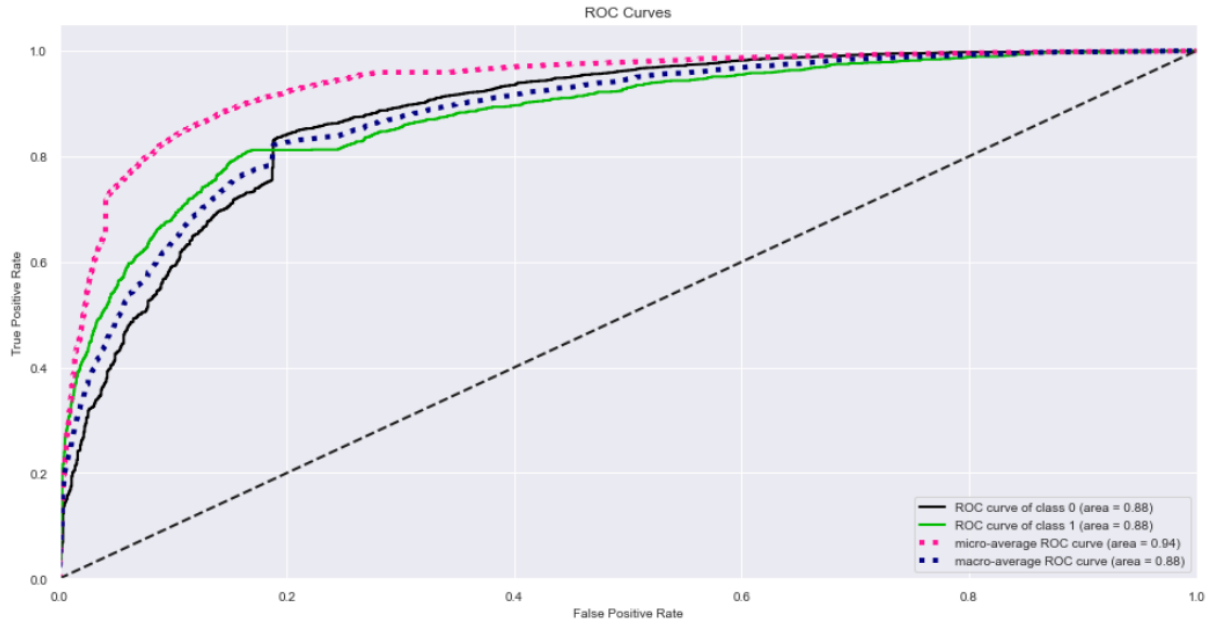
Model Accuracy score: KNN 0.826970227670753

Model Evaluation:

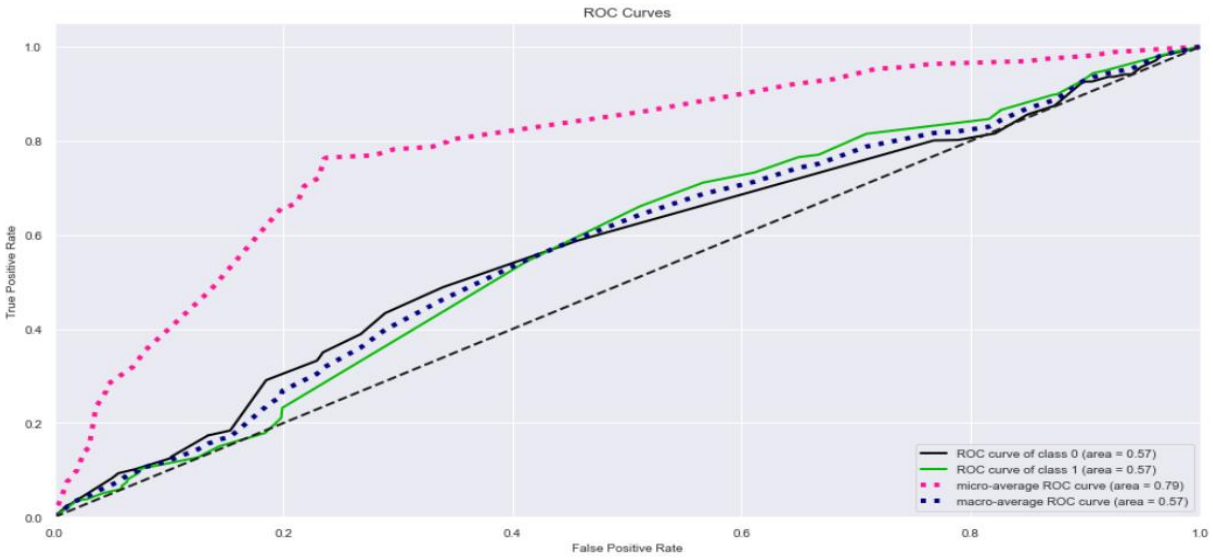
Logistic Regression:



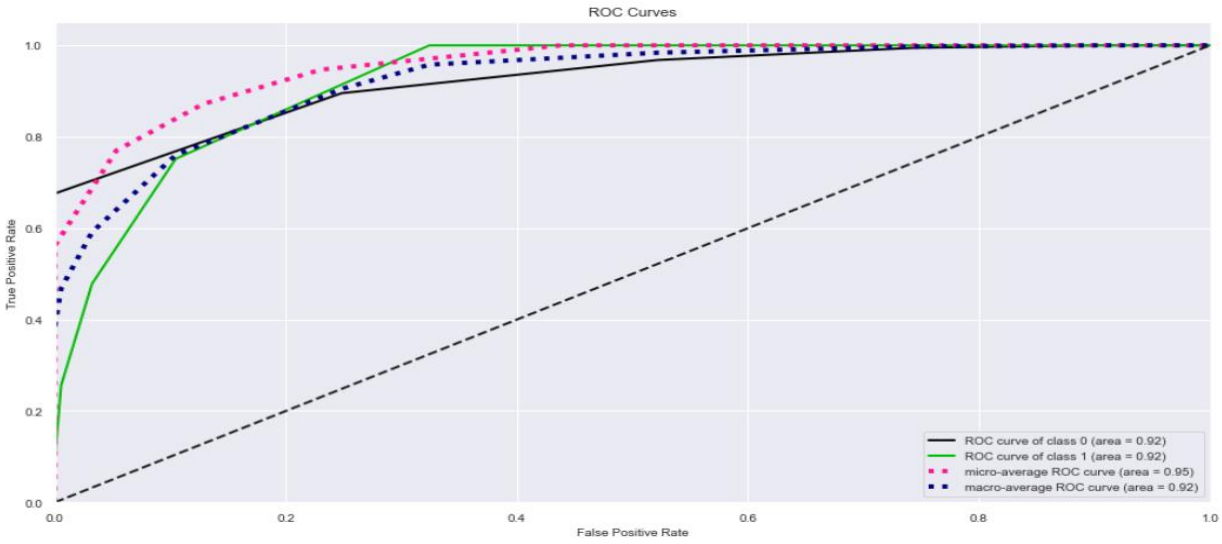
SVC:



Random Forest:



KNN:



Final results:

	MODEL	ACCURACY_SCORE
1	Random Forest Classifier	0.857443
2	SVM Classifier	0.853940
3	Logistic Regression	0.829422
4	KNN Classifier	0.826970

Assumptions and Limitations

Assuming a churn model is good at its features compared to other machine learning models. One can accomplish a churn model if and only if the data scientist is sound in domain knowledge. And also, skill and creativity are needed to construct a robust feature set with information predictive of a churn event. Several roadblocks can arise at this stage, such as target leakage, unavailable or missing information, or the need for optimal feature transformations.

The main limitation observed with the use case is constructing the target variable for the churn event. This process may not always be straightforward in all use cases as it depends on the dataset we are dealing with. For example, in a banking environment where customers cancel and renew frequently, how can we define churn? What about a scenario where customers can subscribe to multiple bank schemes/products?

In-depth, EDA can reveal irregularities, correlations, outliers, and relationships that domain knowledge alone wouldn't account for. A deep insight of EDA and building a robust ML often have to occur before we embark on building an overall churn model.

Challenges

- Anticipating whitespaces in data and need to work on data alignment.
- Incorrect variable types.
- Python package-related issues.
- It is challenging to measure churn – mainly when we try to measure it based on historical data. The philosophy "The future might resemble the past" – however, we need to know that nothing is certain. Below are the challenging factors that need to be accomplished to solve the business problem.

1. Finding a pattern
2. Finding the root cause of churn

- I am relying on the Kaggle dataset as this is usually clean, and I anticipate no missing or insufficient data that needs to be substituted with dummy data.
- Since the data is not current, Modeling and forecasting may not apply to the present-day scenarios.

Recommendations and use cases:

Customer churn analytics use customers' transactional data to predict the root cause of churns. For better results, this Model needs to be integrated with the company's existing CRM or support systems. It exemplifies the number of customers we have lost or are planning to exit.

The churn analytics enables organizations to track customers' subscribing events to show the customer journey patterns. With the help of ML techniques, we can compare the customer behavior w.r.t retained customers and analyze what went wrong.

Below are the few other use-cases the churn predictions can be applied.

- Music and video streaming services
- Media.
- Telecom companies.
- Software as a service provider.

Ethical Assessment

I would seek customers' consent before using the customer data in predictive modeling applications that drive the marketing strategies. Companies may have to explain how customer data use benefits consumers to obtain consent. Whether the bank uses the customer's transactional data to drive value to add promotions on their schemes or product, improve inventory control, drive research and development, or any other legitimate purpose, the consumers have to evaluate the trade-offs and, in most cases, and be willing to allow the use of their data after the "opt-in" consent is received. I would consider the ethical assessment factors for the given use case below.

- ✓ **Informed consent** often refers to written consent by a customer to participate in any given survey activity where PII data may be collected.
- ✓ **Confidential data** refers to PII that refers to a particular individual but is kept confidential such as medical or service records.
- ✓ While it is essential to have **transparent processes for collecting data**, it is equally crucial to have transparent procedures for sharing data.

Conclusion

Many organizations and companies are suffering from the fact that their customers constantly change their service provider and join the services of rival organizations. Due to many customers and the high speed of data generation in business related to retail customers, there is no possibility of achieving business intelligence without using machine learning mechanisms. Therefore, it is essential to find a suitable prediction model for customer churn using machine learning methods for a bank's business analysis unit. This use case is aimed to implement a churn prediction model using machine learning classifiers.

The following conclusions are depicted from the analysis on the features:

- The customers who used products 3 and 4 stopped working with the bank, and all customers subscribed to product number 4 were exited.
- A credit score below 450 had high abandonment rates.
- Customers aged between 40 and 65 are more likely to quit the bank.
- Predictions were made with a total of 4 classification models. The highest accuracy is observed with the KNN classifier.
- Accuracy scores and ROC metrics were calculated for each Model, and results were displayed.

Appendix

Parameters:

Parameters Description		
Independent Variables		
1	RowNumber	Row number of the customer in the csv file. There were 10,000 customers in total.
2	CustomerId	Unique identification of each customer from the bank's records
3	Surname	Surname of the customer
4	CreditScore	A score the bank assigns to each customer based on the customer's personal credit history to measure the customer's creditworthiness.The higher the credit score, the
5	Geography	The location where the customer lives
6	Gender	Customer's gender (male or female)
7	Age	Customer's age
8	Tenure	How long the customer has been with the bank in years
9	Balance	Present monetary value of a customer's account
10	NumOfProducts	Products the customer is currently using from the bank (e.g., Internet banking, loans, and currency or savings accounts, among others)
11	HasCrCard	Binary indication of whether a customer possesses a credit card or not
12	IsActiveMember	Indicator of whether a customer has used any of the bank's products in the last 6
13	EstimatedSalary	Estimated customer salary
Dependent Variable		
14	Exited	This depends on a variable that indicates if the customer has left the bank after 6

Logistic regression: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

SVM: https://scikit-learn.org/stable/modules/svm.html#:~:text=Support%20vector%20machines%20(SVMs)%20are,than%20the%20number%20of%20samples.

KNN: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Random Forest(RFC): https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

roc_curve: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html

PII – Personal identifialble information

References:

Kaemingk, D. (2018, August 29). *Reducing customer churn for banks and financial institutions*. Qualtrics. Retrieved January 14, 2022, from https://www.qualtrics.com/blog/customer:churn:banking/

Tausend, F. (2018, September 19). *Eliminating churn is growth hacking 2.0*. Medium. Retrieved January 14, 2022, from https://blog.markgrowth.com/eliminating:churn:is:growth:hacking:2:0:47a380194a06

Tausend, F. (2019, January 15). *Hands:on: Predict customer churn*. Medium. Retrieved January 14, 2022, from https://towardsdatascience.com/hands:on:predict:customer:churn:5c2a42806266

Wuraolaifeoluwa, . (2021, May 21). *Prediction of customer churn in a bank using machine learning*. Medium. Retrieved January 14, 2022, from https://wuraolaifeoluwa.medium.com/prediction:of:customer:churn:in:a:bank:using:machine:learning:5a456c184ed1

Álvarez, C. (2018, September 25). *Big Data and privacy: New ethical challenges facing banks: BBVA*. NEWS BBVA. Retrieved January 14, 2022, from https://www.bbva.com/en/big:data:privacy:new:ethical:challenges:facing:banks/

Altexsoft, .. (2019, May). *Customer churn prediction using machine learning: Main approaches and models*. KDnuggets. Retrieved January 16, 2022, from https://www.kdnuggets.com/2019/05/churn-prediction-machine-learning.html

., metarouter. (2021). https://www.metarouter.io/blog-posts/the-ethics-of-collecting-consumer-data [web log]. Retrieved January 30, 2022, from https://www.metarouter.io/blog-posts/the-ethics-of-collecting-consumer-data.

10 Questions:

1. How this Model does help in banking strategies?
2. What do you mean by the term "churn"? Why is this important to predict?
3. How do end-users access the results?
4. From EDA, What is the percentage of churn customers?
5. What are the classifiers used for this use case, and why?
6. What factors impact customer churn, and how do we reduce customer churn rate?
7. What do we infer from class imbalance?
8. From EDA, what is the role of "products owned" that contributes to the churn rate?
9. From EDA, what is the role of "Having a credit card" that contributes to the churn rate?
10. Which classifier is accurate for this use case?