NAME: Wairiuko Samuel

COURSE: Data Science Project.

TITLE: Bank Customer Churn

DATE: May 30, 2020

CONTENTS

ABSTRACT

Customer churn refers to when a customer ceases his relationship with the company. This report is about Bank customer Churn Prediction aimed to analyze the key variables that are associated to Churn, identify the customers at risk and to come up with solutions to minimize attrition. In this report I have discussed the profiles of the banks in three European countries namely France, Germany and Spain. I have prepared this report as a python project requirement for Advanced Diploma in Data Science and Application and focused on bank attrition Exploration and prediction models.

PROBLEM STATEMENT

All banks acknowledge customer churn as a significant business problem. Churn is major threat to profitability, revenue and reputation to banks. The goal of this attrition analysis is to find out customer groups that have high probability to attrite and then the bank can conduct marketing campaign to change the behaviour in a desirable direction. Before attrition occurs, some customers pay slowly their balance until they become inactive and others pay off quickly their debt and then write-off their Accounts. The middle aged also tend to have a high churn rate as they look for better offers or relocate to different areas. Conducting customer churn analysis and building accurate prediction models will help to produce insights that can offer good customer retention strategies. This will help to improve customer royalty, minimize unnecessary marketing cost and boost revenue to the banks.

Objectives

- ✓ Business Problem
- ✓ Hypothesis Generation
- ✓ Exploratory Data Analysis
- ✓ Data Cleaning
- ✓ Feature Engineering
- ✓ Model Building
- ✓ Model Results Analysis
- ✓ Summary (Conclusion & Recommendations)

Hypothesis Generation

- ✓ Customers with high loan balance and low credit score are more likely to churn.
- ✓ The fewer the number of products a customer have in the bank the higher chances of churn.
- ✓ Middle aged customers between 18 and 40 years are more likely to churn than those above 70 years.

Properties of the Dataset

This project is about Bank Customer Attrition in France, Germany and Spain. I got the dataset from www.kaggle.com. The dataset has 14 variables and 10000 observations. The Target variable is "Exited" other variable's includes Row Number, Customer Id, Surname, Credit Score, Geography, Gender, Age, Tenure, Balance, Number of Products, Has Credit Card, Is active Member and Estimated Salary.

Data Preprocessing and Exploration

- ✓ I imported the data
- ✓ The data had no missing values or duplicates.
- ✓ Dropped the following columns; Row Number, Customer Id, Surname, and later on when building the model, I also dropped Geography and Gender and feature Engineered new encoded dummy variables.
- ✓ Checked for outliers using boxplot.

Data Exploration Analysis

- ✓ Attrition analysis shows 80% of customers were active while 20% of the *customers churn.*
- ✓ In terms of number of customers per country France had more customers with 50% followed by Spain 25% and Germany with 25% of the total customers.
- ✓ Most churn was recorded in the group aged between 39-60yrs.

- ✓ The average age was 40yrs in all the three countries, the oldest customer was 91 years old.
- ✓ The average credit score was 650, The customers who had Credit score below 380 exited.
- ✓ The churn rate for Gender was 55% Female and 45% Male.
- ✓ The average estimated salary for Female was 100601 and Male 99664.
- ✓ Majority of customers who exited had 2 or less than 2 products.

FEATURE ENGINEERING

I used One-Hot-Coding to creating dummy variables for geographical variable that has 3 categories – 'France, 'Germany' and 'Spain'. One hot coding removed this variable and generated 3 new dummy variables 0 and 1 using 'get dummies' function of Pandas. I also used Label Encoding to convert Gender column which had Male and Female column to Gender encoded with dummy variables of 0 and 1

MACHINE LEARNING MODELS

The Logistic Regression classifier

- ✓ Given the customer records the model was able to predict attrition where there was attrition correctly with an accuracy of 79% for the train and 78.68% for the test (True Positives)

K-nearest neighbors (KNN)

- ✓ Maximum KNN score on the test data: 84.68%
- ✓ Given the customer records the model was able to predict attrition where there was attrition correctly with an accuracy of 90.8% for the train and 87.12% for the test. (True Positives)

The Random Forest classifier

- ✓ Given the customer records the model was able to predict attrition where there was attrition correctly with an accuracy of 90.85% for the train and 87.12% for the test. (True Positives)

Model Accuracy Summary

Before hyperparameter Tuning

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	80.94	80.68
1	K-nearest neighbors	87.02	82.88
2	Random Forest Classifier	100.00	86.88
3	Support Vector Machine	86.82	86.28
4	Decision Tree Classifier	100.00	79.88

Model Accuracy Summary

Model Accuracy After Hyperparameter Tuning

	Model	Training Accuracy %	Testing Accuracy %
0	Tuned Logistic Regression	79.02	78.68
1	Tuned K-nearest neighbors	81.66	76.24

	Model	Training Accuracy %	Testing Accuracy %
2	Tuned Support Vector Machine	100.00	79.64
3	Tuned Decision Tree Classifier	81.13	81.52
4	Tuned Random Forest Classifier	90.85	87.12

Summary Findings and Conclusion

✓ There exists a strong relationship between the number of products a customer has and the churn rate i.e. the higher the number of products the lower the rate of Churn. Majority of customers who exited had 2 or less than 2 products

✓ The results show most churn is in the group aged between 39-60yrs. Probably because they keep looking for better offers or relocate to different areas.

✓ Tuned Random Forest Classifier was the best model for prediction with training accuracy of 90.85% and test accuracy of 87.12% followed by Tuned Decision Tree Classifier with training accuracy of 81.13 and test accuracy of 81.52% in predicting whether the customer exited or not.(The difference between train and test is not more than 5% so it is a balanced model good for prediction that is not overfitted or underfitted)

Recommendations

✓ To improve the performance of the model accuracy to 98% without overfitting the model.

✓ Random Forest is a better compared to the other models

✓ Prioritize the most valuable of at-risk customers aged between 39-60 years through offering them better rates to improve the customer royalty.

✓ The banks should Cross-Sell more than two products to the existing customers to boost bank revenue and reduce the marketing expenditure of acquiring new customers.

Acknowledgements

1. **Machine learning, Data Mining and Big Data Analytics Lecture Notes by Gitimoni Saikia**
2. **Python Project Lecture Notes by Vijay Kumar**