BANK CUSTOMER CHURN PREDICTION MODEL

Submitted as a part of DATA SCIENCE AND BIG DATA ANALYTICS
Course Requirement

By

Asha Gutlapalli
RA1611003010352

Priyam Dutta
RA1611003011211

Under the supervision of

Mrs. B. Gracelin Sheena
Teaching Associate
Department of Computer Science and Engineering SRM Institute of Science and Technology Kattankulathur, Chennai

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

Asha Gutlapalli and Priyam Dutta studying IV year B.Tech in Computer Science and Engineering at SRM Institute of Science and Technology, Kattankulathur, Chennai, hereby declare that this Mini project is an original work and we have not verbatim copied *or* duplicated any material from sources like internet or from print media, excepting some vital company information *I* statistics and data that is provided by the Technical organizations itself

Signature of the Student:

Date:

Place:

ACKNOWLEDGEMENT

It is a matter of great pleasure and privilege to present this report for our Data Science and Analytics Mini Project. Through this report, we would like to thank the people whose consistent support and guidance has been the standing pillar in the architecture of this report.

We would like to express our gratitude towards **Mrs. B. Gracelin Sheena,** our faculty for Data Science and Big Data Analytics, who gave us valuable feedback and guidance throughout the development of the project. She pushed us consistently and followed up regularly, which kept us on track. Her encouragement and support is what made the project what it is.

ABSTRACT

Companies that shift their focus to customer retention often find it to be a more efficient process because they are marketing to customers who already have expressed an interest in the products and are engaged with the brand, making it easier to capitalize on their experiences with the company. In fact, retention is a more sustainable business model that is a key to sustainable growth. Customer retention refers to the activities and actions companies and organizations take to reduce the number of customer defections. The goal of customer retention programs is to help companies retain as many customers as possible, often through customer loyalty and brand loyalty initiatives. In this paper, we develop a model for achieving customer retention in order to help companies to enhance their approach of catering to their customers. Data is pre-processed and normalized before the input is given to the model to avoid inconsistent data which improves the efficiency. Machine learning concepts are used for implementing supervised learning techniques which train on labeled data and make predictions on new unseen data. Classification algorithms are utilized to categorize the customers in the company into either retainable or not retainable based on the profiles of their users. Artificial neural networks are employed to make the model with the aid of weights and features extracted. Data Visualization techniques have been instrumental in representing the results. The model produces a high accuracy and performance and also reduces discrepancies in the data.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Customer churn, also known as customer attrition, occurs when customers stop doing business with a company. The companies are interested in identifying segments of these customers because the price for acquiring a new customer is usually higher than retaining the old one. For example, if Netflix knew a segment of customers who were at risk of churning they could proactively engage them with special offers instead of simply losing them.

In the world of service based industries, determining the audience is of utmost importance. The profits they make depend directly on the customers they have. As such it is important for any service provider to have a steady customer base.

This could not be any truer in the banking sector. With a multitude of banks coming up, each of them has to identify their most valuable customers to not only ply their trade but also cut out the competition. With a rise in the dissemination of knowledge due to the employment of technological tools, the customers in this sector are becoming more conscious and aware than ever before. So it is imperative for every bank to find their own niche which is the audience they cater to, while also reeling in as many new clients as possible.

Customer churn in an important factor in financial services. We have identified that the industry's performance metric is the amount of customers they have. As such, a system that could potentially identify customers that are most likely to stop doing business with the bank would be very beneficial. Identifying such customers would allow banks to put in efforts towards specific customer retention and could help them improve their services in the long run.

Using a consolidated dataset which has features such as customer's balance, activity, assets and quitting status, we can apply a multitude of learning algorithms to create a customer churn predictor which would help identify the target clientele for retention measures. This is the crux of the project and an elaborate set of data analysis tools have been applied to realize this goal.

CHAPTER 2

LITERATURE REVIEW

In the paper “Churn Modes for prepaid customers in the cellular telecommunication industry using large data marts”(2010), Martin Owczarczuk tested the usefulness of popular data mining models to predict churn of the clients of the Polish cellular telecommunication company. It deals with prepaid clients who are more likely to churn, less stable and have lesser information. The feature set was derived from usage and had 1381 features. The stability of models across time was tested for all the percentiles of the lift curve. It was found that linear models were a better choice than decision trees based on stability.

The paper “Churn detection via customer profile modeling” (2007) discussed the usage of customer profile modeling in aiding service providers in analyzing customer behavior, designing customized service plans, and preventing churn activities. It proposed using a functional mixture model to profile customer behavior in order to identify and capture churn activity patterns.

The paper “Handling class imbalance in customer churn prediction” (2009) looked into the importance of class imbalances when sampling churn datasets. It used more appropriate evaluation metrics such as AUC, lift, etc. It also investigated the increase in performance of sampling and modeling techniques (gradient boosting and weighted random forests). Oversampling turned out to give improved prediction accuracy, especially when evaluated with AUC.

The paper “Predicting credit card customer churn in banks using data mining”(2008) by V. Ravi and Dudyala Anil Kumar explored churn prediction using data mining. It developed an ensemble system incorporating majority voting and involving multilayer perceptron, logistic regression, decision trees, random forests, radical basis function and support vector machine. Undersampling, oversampling and a combination of the two was used to balance the skewed dataset.

“Modeling and predicting customer churn from an insurance company” (2011) presented a dynamic modeling approach for predicting individual customers’ churn likelihood. A logistic longitudinal model that incorporated time dynamic explanatory variables and interactions was fitted to the data. The paper also discussed applying generalized additive models to identify the clients likely to leave the insurance company.

CHAPTER 3

METHODOLOGY

Data Reading and Row Extraction

We first read the data using the pandas library, using the CSV reader method. This stores the data in the form of a dataframe. The dataframe can be used for data exploration and other analytical purposes. We then perform row extraction for non-numeric indices to build our feature set upon which the model can be built.

Label Encoding

Label encoding refers to converting the labels into numeric form so as to convert it into machine-readable form. This helps in the implementation of machine learning algorithms. This converts our attributes into a numeric range of values.

One Hot Encoding

Sometimes in datasets, we encounter columns that contain numbers of no specific order of preference. The data in the column usually denotes a category or a value of the category and also when the data in the column is label encoded. To avoid confusing the machine learning algorithm the data in this column should be one hot encoded. This refers to splitting the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.

Standard Scaler

This is a preprocessing tool used to standardize features by removing mean and scaling to unit variance. This centers the dataset and saves it as an internal object state, which is then fitted to parameters on the training set.

Keras.Sequential

Keras is a machine learning toolkit package for python. The Sequential model is present in Keras and is a linear stack of layers. We first specify the input shape which is defined by parameters such as “Dense”, ”batch_size” and so on. After this, the compilation phase begins. Here the model’s learning process is configured. The compile() function uses an optimizer, a loss function and a list of metrics. These arguments help determine the target set and allow us to tweak the metrics.

Training

Using the numpy package's arrays of input data and labels, Keras models implement the fit function to train the predictive model.

Visualization

We describe the performance of the model on our test set by using a confusion matrix. This allows easy identification of confusion between classes (here, users likely to churn or not). The number of correct and incorrect predictions are summarized with count values and broken down by each class. This gives an insight not only to the errors being made by a classifier but more importantly the types of errors that are being made.

CHAPTER 4

DATA ANALYSIS

The Dataset

The Churn_Modelling.csv dataset was obtained from the online repository Kaggle. This data set 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.

Analysis

We use the pandas, numpy and matplotlib libraries for our exploration and analysis.

We ascertain the structure of the dataset, i.e., the features, value ranges, etc.

```
In [4]: dataset.head()
```

Out[4]:

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalar
0	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.8
1	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.5
2	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.5
3	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.6
4	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1

In [5]:
```
dataset.info()
```

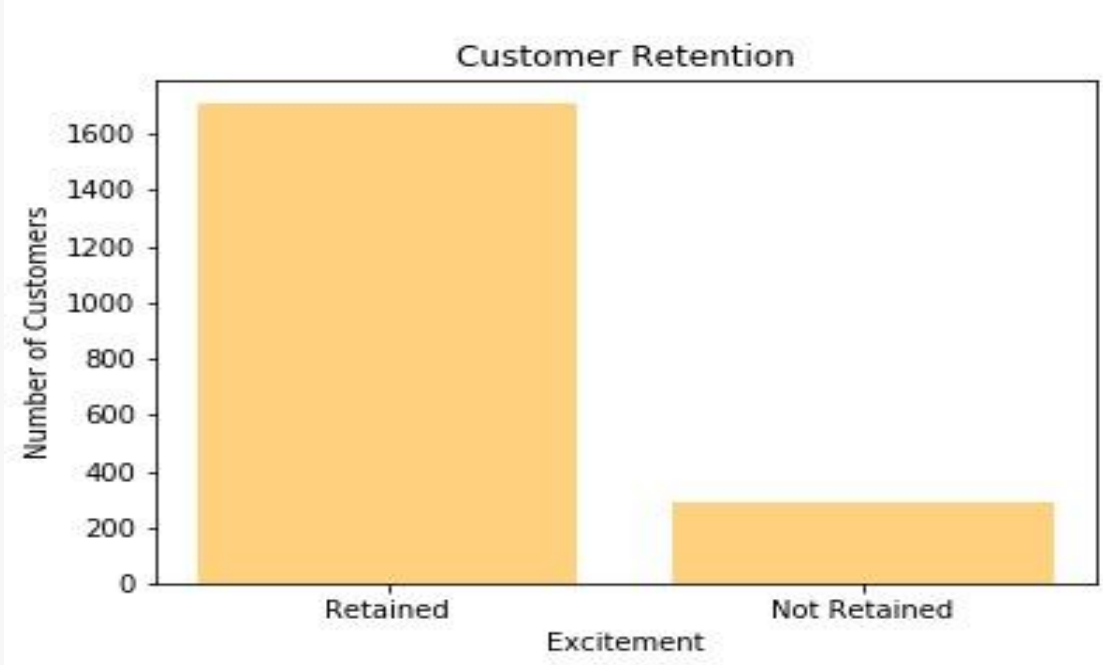
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10000 entries, 0 to 9999
Data columns (total 14 columns):
RowNumber          10000 non-null int64
CustomerId         10000 non-null int64
Surname            10000 non-null object
CreditScore        10000 non-null int64
Geography          10000 non-null object
Gender             10000 non-null object
Age                10000 non-null int64
Tenure             10000 non-null int64
Balance            10000 non-null float64
NumOfProducts      10000 non-null int64
HasCrCard          10000 non-null int64
IsActiveMember     10000 non-null int64
EstimatedSalary    10000 non-null float64
Exited             10000 non-null int64
dtypes: float64(2), int64(9), object(3)
memory usage: 976.6+ KB
```

In [7]:
```
dataset.dtypes
```

Out[7]:
```
RowNumber            int64
CustomerId           int64
Surname             object
CreditScore          int64
Geography           object
Gender              object
Age                  int64
Tenure               int64
Balance            float64
NumOfProducts        int64
HasCrCard            int64
IsActiveMember       int64
EstimatedSalary    float64
Exited               int64
dtype: object
```

CHAPTER 5

RESULTS



We can infer from the above test results that 85% of the customers in the company are predicted to be retained. But remaining 15% of the firm's customers are predicted to not retain. The model trained gives high accuracy of about 90%. The accuracy on the test data is estimated as 86%. The trade- off between bias and variance is a good for making forecasts on unseen new data. The evaluation parameters used for assessing the model are precision, recall and F1 score that is summarized in a confusion matrix. The confusion matrix consists of true positives, true negatives, false negatives and false positives. The results are finally visualized using Data visualization techniques like bar graph which consists of the predicted number of customers retained and not retained.

CHAPTER 6

CONCLUSION

We can conclude that the model developed for Data Science and Big Data Analytics Mini Project has the capacity to demonstrate data science and machine learning concepts. The data pre-processing steps are carried out to make sure that the data is consistent and normalized before giving it as input. The model is created in a supervised machine learning approach. Classification algorithm is employed to categorize the instances of the dataset into either retained or not retained. After sufficient training, the model is made to foretell whether the customer can be with held in the firm. With the help of this information, companies have the ability to understand what kinds of profiles maintained by the customers prefer the company and what kind of profiles do not desire to invest in the company. This information is crucial as this can aid them in catering to the users that do not show interest. Eventually increasing the profits and the performance of the industry. The model achieves high accuracy and performance. Moreover it can be used by the industries that relate to this domain like employee attrition and many more. It has a wide scope for applications. Data visualization of the results is also done to get a better understanding of how the customers are distributed with different backgrounds and accordingly how the predications are made. This project can be beneficial to companies that aspire to understand their users better and improve their experience as well as the firm’s performance.

BIBLIOGRAPHY

1. Clara-Cecilie Gunther, IngunnFrideTvete, KjerstiAas, GeirIngeSandnes, OrnulfBorgan,"Modelling and predicting customer churn from an insurance company"*Taylor and Francis*(2011)

2. Dudyala Anil Kaamar, V. Ravi, “Predicting credit card customer churn in banks using data mining” *Int. J. Data Analysis Techiques and Strategies, Vol. 1, No. 1*(2008)

3. J. Burez, D. Van den Poel “Handling class imbalance in customer hurn prediction” Elsevier (2009)

4. MarcinOwczarczuk “Churn Models for prepaid customers in the cellular telecommunication industry using large data marts” Elsevier(2010)

5. ZhiguangQian, Wei Jiang, Kwok-Leung Tsui “Churn detection via customer profile modelling” Taylor and Francis (2007)