Customer Churn Prediction

Prof. Swati Shriyal ; Students: SARISH CHACHARE, DAKSH KAMBLE, ANUSH JAIN

Department of Engineering, Sciences and Technology

Vishwakarma University, Pune - 411037, Maharashtra, India

***Abstract – Knowing the client churn rate could be a key indicator for any business. in step with a study by Bain & Company, rising the customer retention rate for existing customers by simply five p.c will improve a company’s profit by twenty-five to 95. Because it is very important for each company to carry their customer as their profit directly depend upon customers. Thus, it's important to analyse the purchasers who are seemingly to depart bank.***

***Keywords- Python, Machine Learning, Data Science***

***­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­­***

I. INTRODUCTION

Predicting churn means detecting which customers might abandon a service or cancel a service subscription. This is a critical forecast for many businesses, as acquiring new customers often costs more than retaining existing customers. Once you can identify which customers are at risk of being cancelled, you need to know exactly what marketing action to take for each individual customer to maximize the chances of them staying. Different customers show different behaviours and preferences, so they cancel their subscriptions for various reasons. It is therefore essential to communicate proactively with each of them to keep them in your list of customers. You need to know what marketing action will be most effective for each customer and when it will be most effective. Customer churn is a common problem for all businesses in many industries. If you want to grow your business, you need to invest in acquiring new customers. Each time a customer leaves, it represents a significant lost investment. Time and effort must be channelled into replacing them. Being able to predict when a customer is likely to go and offer their incentives to stay, she can offer huge savings to a business.

II. PROPOSED SYSTEM

* Identify and visualize which factors contribute to customer churn:
* Predicting the churn rate for a customer and classify them by learning about different classification algorithms.
* Comparing and evaluating different algorithms based on its performance.

III. METHODOLOGY

The main steps in model building are:

1. Data Collection

Data collection is the process of collecting and measuring information about targeted variables in an established system, which then allows you to answer relevant questions and evaluate the results. Data collection is a component of research in all fields of study, including the physical and social sciences, the humanities, and business. Although methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same. The goal of all data collection is to acquire quality evidence that allows analysis to lead to the formulation of compelling and credible answers to questions that have been posed. Here, the dataset comes from Kaggle, which is a free repository for all data scientists. The features or variables in dataset are:

* customer\_id, unused variable.
* credit\_score, used as input.
* country, used as input.
* gender, used as input.
* age, used as input.
* tenure, used as input.
* balance, used as input.
* products\_number, used as input.
* credit\_card, used as input.
* active\_member, used as input.
* estimated\_salary, used as input.
* exited used as the target. 1 if the client has left the bank during some period or 0 if he/she has not.

1. Data Pre-processing

Data pre-processing is the process of transforming raw data into an understandable format. It is also an important step in data mining because we cannot work with raw data. Data quality should be checked before applying machine learning or data mining algorithms. In this project, Pandas Python library was used to perform data pre-processing. Attributes that did not affect the forecast were removed from the dataset. Various functions have been used to analyse and study the dataset.

1. Data analysis and visualization

Data visualization is the process of translating large sets of data and metrics into charts, graphs, and other visual elements. The resulting visual representation of the data makes it easy to identify and share trends, outliers, and new insights into real-time information about the insights represented in the data. Here Matplotlib and Seaborn library was used to plot different charts and graphs. Bar graphs were used to observe credit card customers and active members, and the number of products versus outgoing customers was also plotted. Pie charts have been drawn to explore the proportions of target variable classes and 79.63% of the data is for returning customers while the remaining 20.37 are included for outgoing customers. Many pie charts and box plots have been drawn to analyse different situations.

1. Training & Testing with algorithm

A training model is a set of data used to train an ML algorithm. It consists of the sample output data and the corresponding input data sets that influence the output. The learning model is used to run the input data into the algorithm to correlate the processed output with the sample output. The result of this correlation is used to modify the model. This iterative process is called "model fitting". The accuracy of the training dataset or the validation dataset is critical to the accuracy of the model.

In machine learning, model testing is brought up because the method wherever the performance of a completely trained model is evaluated on a testing set. The testing set consisting of a group of testing samples ought to be separated from each coaching and validation sets, however it ought to follow constant likelihood distribution because the coaching set.

IV. ALGORITHM USED

1. Logistic Regression

Regression is one of the statistical processes for estimating how variables relate to each other. It includes a large number of techniques for establishing the model by analysing several variables, when the epicentre of importance is on the shared link between a dependent variable and one or more independent variables. Considering customer abandonment, regression analysis is not widely used because linear regression models are useful for predicting continuous values. But Logit (LR) logistic regression analysis is a probabilistic statistical classification model. It is also used for binary classification or binary forecasting of a categorical value (e.g., real estate rate forecast, customer abandonment) that depends on one or more parameters (e.g., dwelling characteristics, customer characteristics). To solve the complex problem of predicting customer abandonment, the data must first be transformed into an adequate data transformation from the initial data to achieve good performance and sometimes work as decision trees.

2. Random Forest

It works on the divide and conquer approach. It is based on the random subspace method. In this method, multiple trees are trained, and each decision tree is trained by selecting any random sample of attributes from the set of predictor attributes. Each tree matures to its maximum extent depending on the attributes or parameters present. The final decision tree is formed for the forecast based primarily on weighted averages. It can handle thousands of input parameters without deletion. It can also handle missing values ​​in the data set to train the predictive model.

V. FLOWCHART

VI. LIMITATIONS

There are few limitations for my Project. As I am making it on beginners’ level, the accuracy and precision of prediction is quite less compared with existing model. It can further be modified as data cleaning which is carried out was just basic it can have complex cleaning and then the model will have accuracy more than this one.

.

VII. CONCLUSION

The proposed model works and predicts the correct outcome with an accuracy of 78.50 in case of logistic regression and 85.66 in case of random forest. So, from this I can say that the random forest approach will give better forecasting results. Churn prediction modelling techniques attempt to understand the specific customer behaviours and attributes that signal the risk and timing of customer churn. Putting it all together, it's important to anticipate customer abandonment. Effective steps can be taken to build customer loyalty before it's too late. The ability to predict that a customer is at high risk of churn while there is still time to do something is a huge source of additional potential revenue for any online business.

VIII. REFERENCES

1. Lalwani, Praveen & Mishra, Manas & Chadha, Jasroop & Sethi, Pratyush. (2021). Customer churn prediction system: a machine learning approach. Computing. 1-24. 10.1007/s00607-021-00908-y.
2. Rodan A, Faris H, Alsakran J, Al-Kadi O (2014) A support vector machine approach for churn pre-diction in telecom industry. International journal on information 17(8):3961–3970
3. Shaaban E, Helmy Y, Khedr A, Nasr M (2012) A proposed churn prediction model. InternationalJournal of Engineering Research and Applications 2(4):693–697
4. Sharma H, Kumar S (2016) A survey on decision tree algorithms of classiﬁcation in data mining.International Journal of Science and Research (IJSR) 5(4):2094–2097