COURSEWORK TITLE:

Individual Portfolio

Module Code and Title:

7143 CEM – Programming for Data Science

Cohort:

January 2023

Student ID:

13475101

Lecturer:

Mark Johnston / Long Chen

ADD WORD COUNT AND LINK OF THE DATASET

**ABSTRACT**

The telecommunication market is very well developed and due to its rapid rise and development, companies are inclining towards expansion of the subscriber base. There is a need to solve the urgent problem of retaining the customers and predicting the outflow of the customer base in order to survive in the competitive environment. The cost of acquiring customer is more than to retain the existing customers therefore it is necessary to collect information so as to predict customers who are more likely to churn. This reports looks at the usefulness of different machine learning algorithms in predicting the customer churn. The reports focusses on comparing different approaches and methods for customer churn prediction such as logistic regression, decision trees, random forest and gradient boosting, taking into account their ability to tackle imbalanced datasets, identify patterns and produce interpretable predictions. Accuracy, Precision, Recall, F1 score and area under the receiver operating characteristic curve (AUC-ROC) are among the evaluation measures used to ensure a thorough assessment of model performance.

**INTRODUCTION**

Due to fierce competition between companies in the telecommunication industry, customer churn is inevitable. Customer churn is basically when a customer chooses to end the subscription of one company and opts for the services of another company. Customers are very important assets in any business and due to the direct effects on the revenues of the companies, they must find ways to reduce customer churn in order to maintain market position. It is important to understand the customers better and retain every valuable customer (<https://ieeexplore.ieee.org/document/9760309>). Customer relations and customer’s knowledge have positive influence on company’s performance. A survey shows that average annual churn rate stretches between 20% to 40% whereas the cost of retaining the existing customers is 5-10 times much lesser than acquiring new customers. (<https://upcommons.upc.edu/bitstream/handle/2117/104247/DGarcia_etal_resubmit-2016-5-3-v2.pdf;jsessionid=FDBB1B8D710E1DD2E08A2F6714B1EA16?sequence=3>). The profit increases from 25% to 85% when the churn rate decreases by even 5% (<https://ieeexplore.ieee.org/abstract/document/8365230>). This proves that understanding the reasons for customer outflows and customer retention strategy plays a crucial role.

Companies needs to take proactive steps by improving their services, experiences and offering tailored and better rewards for better customer retention. Improving customer’s engagement with the company and reducing customer churn is relevant and important therefore there is an increasing dependency on modern data mining tools, methods, algorithms and machine learning technologies which generate additional opportunities (). As a result of growth in data mining technologies and vast amount of customer data available, companies have started to use data science tools for taking proactive steps in order to understand customers better such as their behaviour and habits (). By using data analysis and precision marketing, a more effective marking approach can be created to retain customers by accurately predicting their churn behaviour, identifying current problems and finding the right recommendations to improve the business of the company.

To address the customer churn issue, in this report, various machine learning algorithms are used in order to build a churn prediction model. Various analysis was also carried out to identify the reasons for customer churn and will suggest ways to reduce the event.

Since the report is part of the coursework Big Data Analytics and Data Visualization module, it’s supposed to show how big data analytics concepts work and how insights can be derived from the data. In addition, the report also covers the installation and set-up of programs such as Tableau (for Data Visualization), PySpark(for Big Data Analytics) and Jupyter Notebook for coding aspect of the report.

**LITERATURE REVIEW**

Many different machine learning approaches and data mining techniques were used to predict the customer churn in telecom companies. Most of the related work either focusses on applying just one method to derive insights for predicting the churn while the other compare different methods.

He et al. (<https://ieeexplore.ieee.org/document/5358641>) developed a Neural Network Algorithm-based model for prediction to address the issue of customer churn of a Chinese Telecom company and achieved an accuracy rate of 91.1%. According to (<https://thesai.org/Downloads/Volume2No2/Paper%204-Churn%20Prediction%20in%20Telecommunication.pdf>), largest sources of revenue loss was because of customer churn which caused an unnecessary burden on telecom companies therefore they developed a decision support system using data mining technique to predict the customer churn behaviour well in advance.

(<https://www.researchgate.net/publication/236625937_A_Proposed_Churn_Prediction_Model>) proposed an easy model using data mining techniques to maintain the records of customer churn and their behaviour. The methods used for classification were Decision Trees, SVM and ANN while K-Means was used for clustering. (<https://doi.org/10.1007/978-981-15-0751-9_9>) suggests using ANN rule-based model to predict customer churn. Implementing this framework will help business develop and deploy intelligent decision support systems. (<http://www.ijstr.org/final-print/feb2020/Customer-Churns-Prediction-In-Telecom-Using-Adaptive-Logitboost-Learning-Approach.pdf>) suggested a prediction model using classification and clustering approaches which identifies customer churn and the reasons behind it by using Adaptive Logitboost. (<https://www.sciencedirect.com/science/article/abs/pii/S1568494620301046>) proposed use of cloud-based ETL(Extract, Transform Load) architecture for data analysis and combining data from different sources. They showed that they were able to identify the exact reason for the churn identify more than 98% of the churners resulting in implementing targeted retention efforts by the sales and support team. I did not find any related work interested in this dataset problem as most of the research papers did not perform the class imbalance techniques. In this report, I used the class imbalance methods to balance the dataset otherwise it can lead to the underperformance of machine learning methods and the issue must be addressed before implementing algorithms.

**DATASET SECTION**

Figure 1 shows the dataset used for this project and report called Telco Customer Churn which is meant to help with focussed customer retention programs. The dataset has 7043 customer records and 21 attributes and has taken inspiration from IBM Samples Data Sets (Blastchar, 2018).

The dataset has information of customer attributes and each row represents a customer which contains information such as if the customer has churned in the last month or not, customer account information, services that customer has subscribed for and demographics information (Blastchar, 2018). Data visualization was carried to further understand the dataset and identify patterns and predict behaviour using Tableau.

DATA CLEANING AND PROCESSING:

Before using any dataset, it is necessary to explore it so as to remove any outlier, missing and null values and changing the attributes into the correct datatypes to improve the performance of the models. We first load the dataset using PySpark as shown in Figure and perform some basic analysis.

* EXPLORING DATA USING PYSPARK:

Figure shows that we have initiated a Spark Session and loaded the dataset. We then proceed to look at the schema of the dataset as shown in Figure . We see that the datatype of TotalCharges column is in the string form when it supposed to be a float. The said column datatype can be changed by running the code displayed in Figure and then we check the dataset schema again to see if the datatype of the column changed shown in Figure. We then try to identify if there any null or missing values in the dataset. Null or missing values can be counted present in each column and needs to be dropped shown in Figure. We can see that there were 11 null values present in the TotalCharges column which we dropped and the updated dataset now has 7032 records. We also drop the Customer ID column since it has no bearing on the outcome of the dataset as shown in Figure.

Figure and Figure uses the shape command to know the number of rows and columns in the dataset and to check the descriptive statistics summary of the dataset such as mean, count, std, etc. respectively. Since the problem is a binary classification problem, we need to make sure that both the classes of the dependent variable is balanced so as to not skew the analysis and avoid biased outcomes. Figure reveal a class imbalance on the dataset so we later use class imbalance techniques so as to balance the dataset.

* EXPLORING DATA USING TABLEAU:

Figure shows that after loading the dataset, a filter was created for the TotalCharges column so as to drop the null values to avoid skewing the analysis.

* + Overall Customer Data Findings (Figure)
    - There are a total of 7043 customers out of which 1869 (26.5%) have already churned.
    - Near equal distribution of male and female customers suggesting that the customers were not found to churn based on gender.
    - Total Revenue generated in a period of 6 years was $16.06M.
    - On average, the monthly tenure of customers is 32 months but the customers with less than 12 months tenure has the highest churn (48.28%).
  + Customer Demographics Findings (Figure)
    - Female customers have a higher churn rate than male customers with 0.5%.
    - Around 74.53% customers who left were not senior citizens indicating the churn rate is higher with younger people.
    - Single customers have a higher churn rate than married customers with 64.21% churn rate.
    - 82.56% customer who left have no dependents.
  + Customer Accounts Findings (Figure)
    - 88.55% customer who left the company had month-to-month contract.
    - Customer who used electronic check payment method had the highest churn rate of 45.29% followed 19.20% customers who used mailed check payment method.
    - 74.91% of the churned customers had opted for paperless billing.
  + Customer Services Findings (Figure)
    - Most customer have signed up for phone services and 69.40% have opted for fibre optic service.
    - 24.56% of the churned customers opted for DSL and streamed movies while 77.37% customers left who had tech ticket but no tech support.

**METHODOLOGY**

OBJECTIVES:

The purpose of this project is to build and compare different machine learning models for customer churn prediction.

The following objectives support this project:

* To Install softwares like Java, Spark, PySpark, Jupyter Notebook and Tableau
* To perform Data Processing and EDA using PySpark and Data Visualization using Tableau
* To select different algorithms for performing data analysis and build models
* To compare the models performance and identify the reasons for the churn.

INSTALLATION and CONFIGURATION:

This section includes the installation of the methods and tools required for this project. PySpark was used to implement the work which acts as an interface to Apache Spark using Python. Tableau and Jupyter Notebook were also installed using MacOS Terminal with 16GB RAM and 87GB of SSD available.

* PySpark Installation:

To use PySpark, we first need to download and install Java, Apache Spark and Python. Jupyter Notebook was also installed for programming environment (<https://sparkbyexamples.com/pyspark/how-to-install-pyspark-on-mac/>)

Figure shows that we need to first install Homebrew which acts like a Missing Package Manager for MacOS to install packages like Java, PySpark. It then asks for the root password which is same password when you log into your Mac for the installation. The successful installation of Homebrew is shown in Figure. We then enter the commands shown in Figure to set the brew to $PATH.

* Java Installation

It is necessary to install Java since it is one of the requirement of PySpark. Figure shows the command to install Java in your system. As shown in Figure, we can see that Java was successfully pre-installed.

* Python and Spark Installation:

We need to install Python since PySpark runs Spark jobs in it hence Figure shows the commands to install Python using Homebrew and we can see in Figure that Python was successfully pre-installed using the latest version. As we know Apache Spark acts as an interface for PySpark, Figure shows the command to run and install Apache Spark in the terminal and we can see that it was successfully installed as shown in Figure. We then run pyspark command in the terminal to launch PySpark shell as seen in Figure. We then proceed to install and launch Jupyter Notebook as shown in Figure and Figure.

* Tableau Installation:

Tableau is a data visualization tool used to analyse and gain insights from the data. Figure shows how the Tableau look once it is launched and that it was successfully installed in my system as shown in Figure.

**EXPERIMENTAL SECTION**

We still needed to perform some more operations on the data to make it suitable. For instance, categorical data needed to be converted into numerical data as machine learning models cannot handle it. Initially, we imported relevant libraries, enabled spark session and loaded the dataset again as shown in Figure. The TotalCharges column was once again converted into float as it was in a string form and 11 null values were identified which were then dropped and so was Customer ID. We can observed that there are 3 possible values for the *MultipleLines* columns which were “No”, “No phone service” and “Yes”. We decided to convert “No phone service” into “No”. The same step was applied to *OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, and StreamingMovies* where “No Internet Service” was converted into “No” as shown in Figure.

Now, we convert all the categorical data into numerical form using *StringIndexer* which were then assembled into one single vector using *VectorAssemble*. The independent features were then stored in a single column called (New Name). The outcome of the dataset were also indexed and then the both the columns(New Columns) were later saved in a new dataframe as shown in Figure. The (New Name) column was then scaled using *StandardScaler* to standardized the data values as shown in Figure. We then split the data into 70-30% for training and testing the models as shown in Figure.

Since this is a classification problem and the outcome is to predict churn or not churn, classification machine learning algorithms such Logistic Regression, Decision Trees, Random Forest and Gradient-boosted Trees were used and metrics such as Accuracy, Recall, Precision, F1-Score and ROC-AUC were used to measure the performance. Initially, we will train our models using actual dataset. Since, the dataset in imbalanced, we then apply class imbalance technique called *Oversampling* to the scaled dataset and apply our machine learning models again and lastly, we will use hyperparameter tuning to increase the performance of the models.

Figure shows the all the 4 classification models code applied using the actual dataset. The codes were referenced from Apache Spark documentation and were customized according to the needs of the dataset (<https://spark.apache.org/docs/latest/ml-classification-regression.html>). Figure shows the process of balancing the dataset using *Oversampling* technique due to which we can see that the data is balanced and then the models were run using the oversampling train and test data sets as shown in Figure. Next, we use few hyperparameter tuning values for each model along with 5-fold cross validation which will give us the best values for each model in improving the performance of the model as shown in Figure. Lastly, the models were again trained using oversample test and train data along with the hyperparameter tuning values as shown in Figure. Finally, we then compare the performance of all the models using *PrettyTable* shown in Fig.

CONCLUSION:

**RESULTS DISCUSSION**

Below mentioned observations were made following exploration and analysis of the data and running the models:

* Customer churn prediction is important as acquiring new customer is costly compared to retaining the existing ones.
* 7043 customers records is less to in accurately predicting the churn since feeding more and more data to models can give better predictions.
* As seen in Figure, customers were less like to churn when stayed with the company compared to new customers.
* Gender did not have any influence on churn and also customers opting for multiple lines and phone services left indicating the multiple lines and phone services were not key influencers for churn.
* Young customers were more likely to churn than senior citizens.
* Company’s tech support can be an influencer since customers with tickets left having no tech support.
* Customers who opted for Fibre optics were more likely to churn than DSL and no internet customers.
* Customer churn was highest who opted for electronic check.
* The top key influencers of customer churn were tenure, month-to-month contract, tech support and fibre optic service.
* Customers were less likely to churn when they had higher total bills.

**CONCLUSION AND RECOMMENDATIONS**

Telecommunication company needs to look after newer customers, especially younger customers since the churn rate is high amongst them and the company also needs to improve their tech services a lot. When monthly charges are less, less customer churn is observed. The dataset had 7043 customers along with their usage behaviour. Customer churn reasons can be useful in accurately predicting the churn. The class imbalance problem was successfully resolved. In the future, the plan is to test the approach on bigger data to predict the churn more accurately. Advanced machine learning models such as XGBoost and Neural Networks along with thorough hyperparameter tuning can be applied to improve the performance of the model and achieving better results. Suggestions from domain expert can also be used in achieving the objectives.

**SOCIAL IMPACT**

* Companies might find new ways through technological advancements to grow their customer base by providing better services and products to their customers.
* Companies can improve overall customer experience by understanding what causes customer churn and problems faced by the customers.
* Companies can use churn analysis to understand where improvements are needs. This can lead to better network coverage, less service disruptions and quick resolutions turnaround time.
* Gaining insights into customer behaviour can lead to fair pricing for the services they receive thereby improving trust levels and decreasing customer dissatisfaction.
* When companies take efforts to reduce churn, it leads to more competition in the market thereby developing new ways to provide better services, ultimately benefitting the customers.
* Companies doing everything to retain customers shows its commitment and support thereby improving the relationship between the organization and the customers.
* Even if the data is publicly available on Kaggle, ethically, the data was anonymized and processed in accordance with the GDPR regulations (<https://ico.org.uk/media/for-organisations/guide-to-thegeneral-data-protection-regulation-gdpr-1-0.pdf>). The dataset is also intended to comply with FAIR guidelines for managing scientific data, encouraging openness, easy to use and reproducibility in research (. <https://www.go-fair.org/fair-principles/>).

**REFERENCES**

**APPENDIX**