CAPSTONE PROJECT

Telco Customer Churn Prediction

Submitted by,

Mangena Suswetha

Abstract

Customer churn is a major problem and one of the most important concerns for large companies. Due to the direct effect on the revenues of the companies, especially in the telecom feld, companies are seeking to develop means to predict potential customer to churn. Therefore, finding factors that increase customer churn is important to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists telecom operators to predict customers who are most likely subject to churn. The model developed in this work uses machine learning techniques on big data platform and builds a new way of features' engineering and selection. In order to measure the performance of the model, the Area Under Curve (AUC) standard measure is adopted, and the AUC value obtained is 93.3%. Another main contribution is to use customer social network in the prediction model by extracting Social Network Analysis (SNA) features. The use of SNA enhanced the performance of the model from 84 to 93.3% against AUC standard. The model was prepared and tested through Spark environment by working on a large data set created by transforming big raw data provided by Telco company. The data set contained all customers' information over 9 months, and was used to train, test, and evaluate the system at Telco. The model experimented six algorithms: Logistic Regression,Neural Network Algorithm, Random Forest, Support Vector Machine, K Nearest Neighbor,Feed Forward Neural Network. However, the best results were obtained by applying Neural Network algorithm. This algorithm was used for classifcation in this churn predictive model.

Keywords: Customer churn, Machine Learning, KNN, Logistic Regression, Random Forest, Support Vector Machine,Neural Network .

Acknowledgements

I am using this opportunity to express my gratitude to everyone who supported me throughout the course of this group project. I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Further, we were fortunate to have Mr.Anbu Joel as my mentor. He has readily shared his immense knowledge in data analytics and guide me in a manner that the outcome resulted in enhancing our data skills.

I certify that the work done by me for conceptualizing and completing this project is original and authentic.

Date: Augest 02, 2022 Name: Mangena Suswetha

Place: Trichy

Certificate of Completion

I hereby certify that the project titled “Telco Customer Churn Prediction” was undertaken and completed (July 2022)

Mentor: Mr. Anbu Joel

Date: Augest 2, 2022

Place – Trichy

Table Of Contents

CHAPTER 1: INTRODUCTION

The telecommunications sector has become one of the main industries in developed countries.The technical progress and the increasing number of operators raised the level of competition [1]. Companies are working hard to survive in this competitive market depending on multiple strategies. Tree main strategies have been proposed to generate more revenues [2]: (1) acquire new customers, (2) up sell the existing customers, and (3) increase the retention period of customers. However, comparing these strategies taking the value of return on investment (RoI) of each into account has shown that the third strategy is the most proftable strategy [2], proves that retaining an existing customer costs much lower than acquiring a new one [3], in addition to being considered much easier than the upselling strategy [4]. To apply the third strategy, companies have to decrease the potential of customer's churn, known as "the customer movement from one provider to another"

1.1 Title & Objective of the study:

1.1.1 Primary objectives

i. To explore the customer churn prediction in telco using machine learning in big data platform

1.1.2 Secondary objectives

i. To investigate the impact of customer churn in telco industry as a whole
ii. To discuss the significance of customer churn models in telco industry
iii. To compare the algorithms that are effective in reducing churn rate in telco Companies

1.2 Need of the Study:

The major contribution of the current study is to present a model of churn prediction that helps telco companies to find users who are more intending to churn. The model formulated in this employs ML algorithms on the platform of Big-Data and develops an innovative way of reducing churn. The significance of this purpose is apparent, given the reality that the expenditure for acquisition of customer is comparatively higher than that of the retention of customer. In consequence, techniques to develop and apply machine learning models are considered necessary and are critical business intelligence applications.The current study is limited to telecom industry only .The study does not use other techniques rather than machine learning techniques .

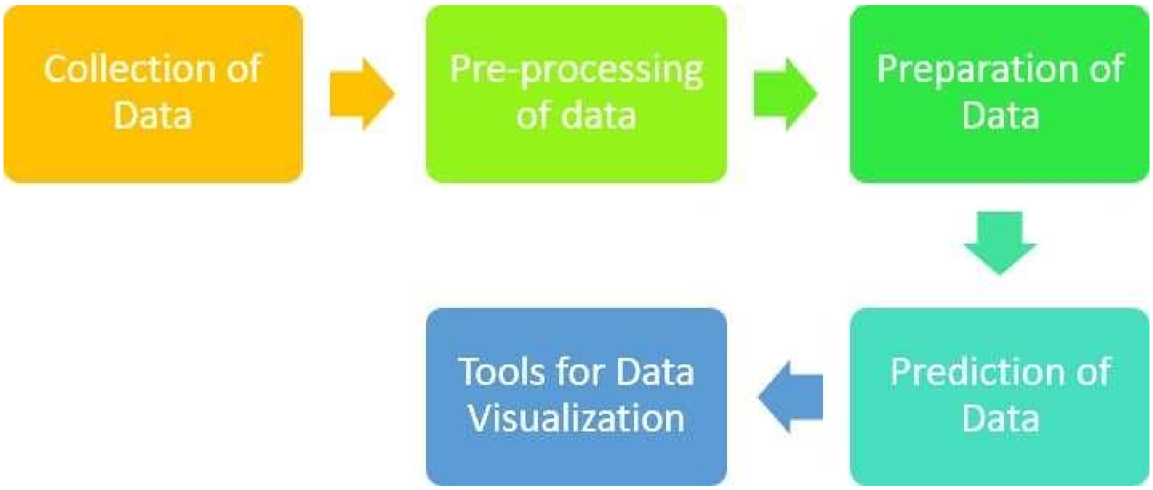
1.3 Business or Enterprise under study:

It is important for Telecom Company to have a churn prediction model in order to prevent their user from moving to another operatorservices. Consequently, the underlying principle of this study is to develop the customer churn prediction model. Machine learning can possibly be the sort of tools which could help telecom companies in churn prediction model. Machine learning is a kind of artificial intelligence tools which give the capability to let computer learns the algorithm instinctively without human contribution. Comparatively, churn prediction in telecom has been considered as unique application domain to churn prediction than othersubscription based industry as a result of the variety, volume and biases of the information. On the basis of the findings, the study noticed that the KNN algorithm surpasses the others with the accurateness for training and testing is the ratio of 80.45% and 97.78% respectively.

The study presents a technique of rule-based decision-making, on the basis of RST (rough set theory), to obtain significant rules of decision linked with non churn and customer churn. The proposed technique efficiently executes categorization of churn from non-churn users, together with prediction of those users who will churn or might likely to churn in the near future. Experiential findings exhibit that rough set theory based on Genetic Algorithm (GA) is the most effective method for obtaining inherent knowledge in decision-based rules form from the publicly accessible, benchmark telecom information. Besides, comparative results exhibit that proposed technique provides a worldwide best solution for churn prediction in the telecom industry, when benchmarked against some high-tech techniques. In the end, the study exhibits that how attribute-level analysis could contribute to develop an effective policy of customer retention that can form an essential part of strategic process of decision-making in the telecom industry.

CHAPTER 2: DATA PREPARATION AND UNDERSTANDING

2.1 Phase I – Data Extraction and Cleaning:



From the above figure 1 the steps used for the proposed system are collection of data, Pre Processing of data, Preparation of data, Prediction of data and Tools for data visualization. The steps are explained below briefly:

2.1.1 Collection of Data

The data that is feasible for analysis in telecommunication data set has been used and the prediction has been carried out for the same.

2.1.2 Pre processing of data

The pre processing of data involves 3 steps namely data cleaning, feature selection and data transformation. Each step is explained below: Data transformation comprises of two explanatory variables which can be transformed from binomial form into binary form to be much applicable for the chosen models.

The data cleaning step involves missing data imputation or handling. Some of the chosen algorithms cannot manage missing data that is why missing value can be transformed by median, mean or zero. However, the replacement of missing data by computed value statistically is a better choice. The used set of data involves missing values in certain numerical variables and two categorical variables. Before training of model, feature selection is one of the most essential factors that can influence the model's performance.

2.1.3 Preparation of data

The main purpose of preparation of data is to improve the quality of data and enhance the performance of data analysis. The preparation of data requires to be undertaken in a much iterative way until a conclusive result is met. The processes of preparation of data involves numerical variables discretization, missing values imputation, selection of feature of most informative variables, transformation from one discrete value set to another and derivation ofnew variables. The process of imputation includes changing the missing
values with whole data based on an estimate from finished values. Making new variables from the information is based on transformation and discretization. Two new variables were formed to estimate the voice and transformation in usage of data. Before the data can be examined the data must be cleaned and keep it prepared so that the desired outputs can be derived from it. Data must be clean so that the errors and redundancy can be eliminated because having such information will lead to improper outcomes as well. In this study a churn examination has been used on telecommunication data here the agenda is to know the feasible consumers that might churn from service provider. The end outcome provides the churn probability for each consumer. To perform the churn examination the logistic regression is used. Logistic regression is a statistical approach where the output variable is categorical rather than continuous. Logistic regression restricts the prediction to be one and zero interval.

2.1.4 Missing values

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of services and others may have something different. In addition, there are some columns related to system configurations and these columns have only null value for all customers.

2.1.5 Prediction of data

The organization is concerned in the final product and it is very essential to indicate their outcome in a graphical representation such a way that is understandable and the output helps organization makes the required predictions which in turns brings profit. There are several components that helps accomplish the same.

2.1.6 Tools for Data Visualization

The best way to acquire the message across is to use the tools of visualization by indicating data visually it is feasible to uncover the essential patterns and the patterns that would be ignored if the statistics alone is considered. In this study Power BI is a component that is used to perfor data visualization. Power BI is a business analytics component which is offered by Microsoft using which reports can be made. In this study the data is cleaned already and the output is populated in a file named prediction which will be helpful to visually display how the data seems and the effect of it.

2.2 Phase II - Feature Engineering

The data was processed to convert it from its raw status into features to be used in machine learning algorithms. This process took the longest time due to the huge numbers of columns. The first idea was to aggregate values of columns per month (average, count, sum, max, min ...) for each numerical column per customer, and the count of distinct values for categorical columns.

Based on the data types and the values, following actions are defined to pre process/engineer the features for machine readability and further analysis:

Columns removed: customerID (not relevent)

2.2.1 Label encoding The following features are categorical and each take on 2 values (mostly yes/no) — therefore are transformed to binary integers

- gender
- Partner
- Dependents
- Churn
- PhoneService
- PaperlessBilling

2.2.2 One-Hot Encoding The following features are categorical, yet not ordinal (no ranking) but take on more than 2 values. For each value, a new variable is created with a binary integer indicating if the value occured in a data entry or not (1 or 0).

- MultipleLines
- InternetService
- OnlineSecurity
- OnlineBackup
- DeviceProtection
- TechSupport
- StreamingTV
- StreamingMovies
- Contract
- PaymentMethod

2.2.3 Min-Max Scaling Values of numerical features are rescaled between a range of 0 and 1. Min-max scaler is the standard approach for scaling. For normally distributed features standard scaler could be used, which scales values around a mean of 0 and a standard deviation of 1. For simplicity we use min-max scaler for all numerical features.

- tenure
- TotalCharges
- MonthlyCharges

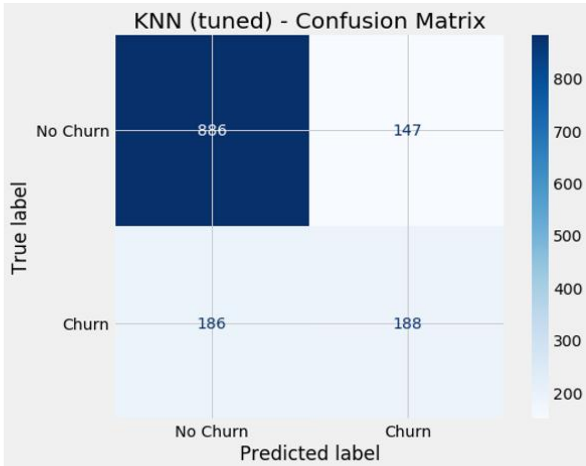
2.3 Exploratory Data Analysis

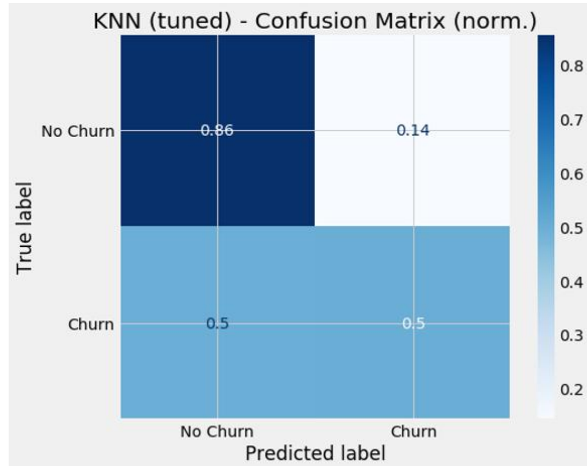
This study uses Kaggle website for data set in predicting and analyzing churn. Kaggle is a site and community for hosting ML competitions. Rivalry ML can be a best way to practice and develop their skills as well as explain their abilities. Kaggle permits users to publish and find sets of data and describe models in a web-based data science surroundings, perform with other scientists of data and ML engineers and enter competition to resolve the barriers of data science. The pre processing steps used for data set are: 1) first the spaces are replaced with values of null in the column of total charges; 2) the values of null are reduced from the column of total charges which comprises 15 percent missing data; 3) then the data is converted to the type of float; 4) after than no internet service is replaced to no for the following columns:Device Protection, Streaming TV, Online Security, Tech Support, Streaming Movies and Online Backup; 5) the values for SenioCitizen is replaced with 0 as No and 1 as Yes; 6) Then the categorical column is made into Tenure; 7) After than the churn and non - churn customers are separated; 8) Finally the numerical and categorical columns are separated.

CHAPTER 3: FITTING MODELS TO DATA

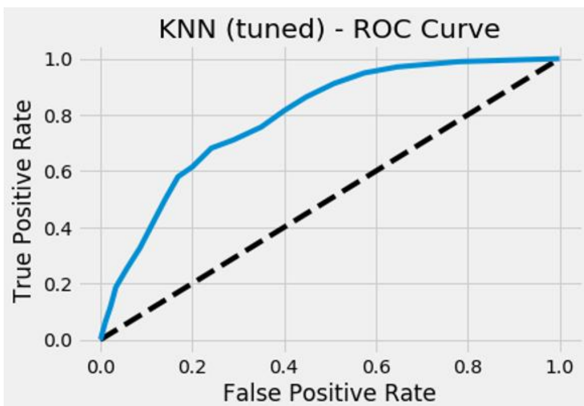
3.1 K-Nearest Neighbor

According to Keramatia et al (2014) K-Nearest Neighbor is one of the most useful and applicable non parametric algorithms of learning. K-Nearest Neighbor is also referred as lazy algorithm that is entire data of training is used at the phase of testing. There is no phase of training and entire points of data are used directly in the testing phase so these entire points required to be employed when it must be tested. K-Nearest Neighbor utilizes the distance between records so as to utilize it for classification. In order to estimate the distance between points K-Nearest Neighbor considers that these points are multidimensional or scalar vectors in feature space. All points of data are vectors of feature space and the label will refer their classes. The easiest case is when the class labels are binary but still it is useful on arbitrary class numbers. In K-Nearest Neighbor one parameter requires to be tuned. K is the number of neighbors/instances that are regarded for instance labeling to some class. The cross validations were carried out using different values of k. K-Nearest Neighbor does not attempt to build an internal structure and computations are not carried out until the time of classification. K-Nearest Neighbor stores only examples of the training information in feature space and the class of an example is decided based on most of the votes from its neighbors. Instance is labelled with class which is much similar among its neighbors. K-Nearest Neighbor decides neighbors based on hamming for categorical variables and distance using Manhattan, Murkowski and Euclidian measures of distance for continuous variables. Estimated distances are employed to recognize training instances set that are nearest to the new point and allot label from these. Despite its simplicity K-nearest neighbor have been used to different kinds of application. For churn K-nearest neighbor is used to examine if a customer churns or not based on features proximity to consumers in every classes .

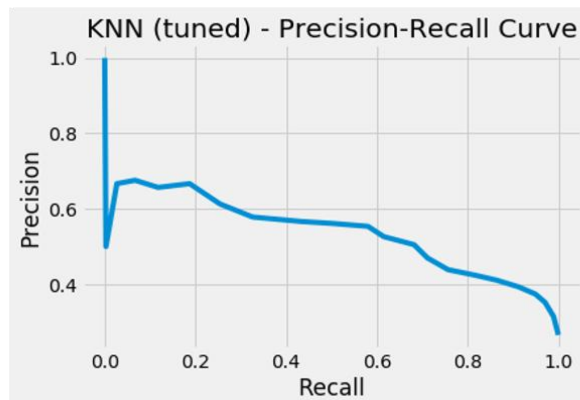





```
Accuracy Score Test: 0.7633262260127932
Accuracy Score Train: 0.8060444444444445 (as comparison)
```




```
AUC Score (ROC): 0.7908718182335858
```

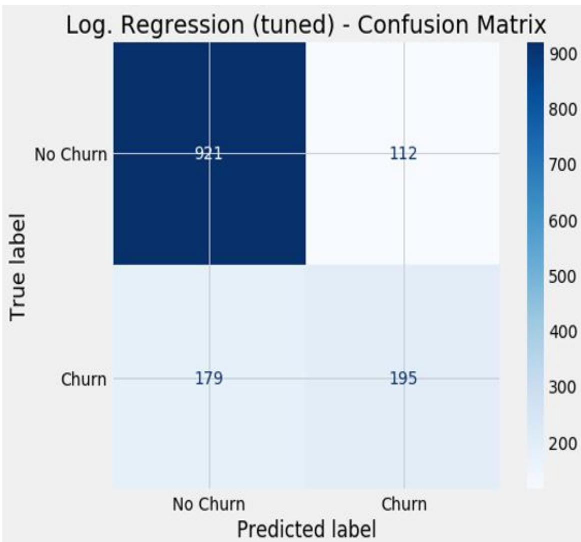


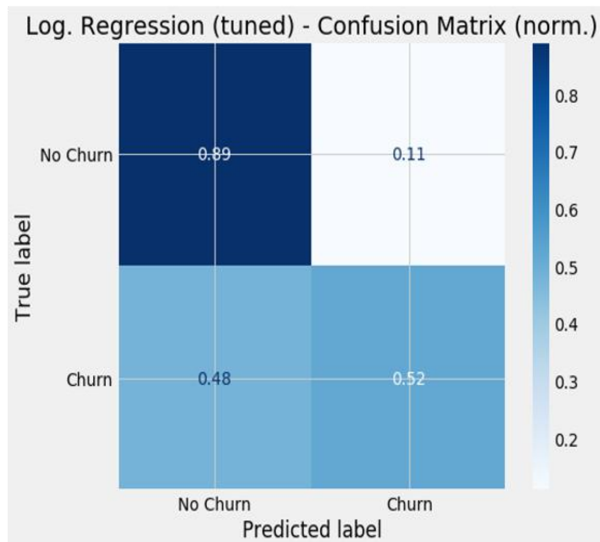

```
F1 Score: 0.530324400564175
AUC Score (PR): 0.5367214270346746
```

3.2 Logistic Regression

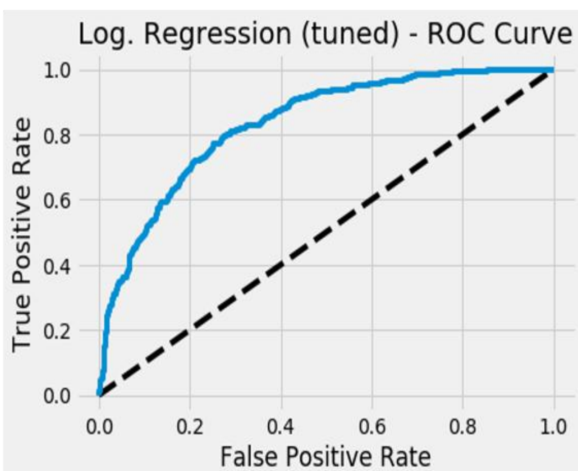
Logistic regression is the proper model of regression analysis to utilize when the dependent variable is binary. Logistic regression is a predictive examination used to describe the relation between an independent variable set and dependent binary variable. For churn of customer logistic regression has been used to estimate the probability of churn as a function of customers characters or variables set (Sahu et al, 2018). According to Hassouna et al (2016) Logistic regression is also used to find the customer churn occurrence probability. Logistic regression is based on a mathematically oriented method to examine the impact of variables on others. Prediction is made by comprising a group of equations linking values of input with the output field.

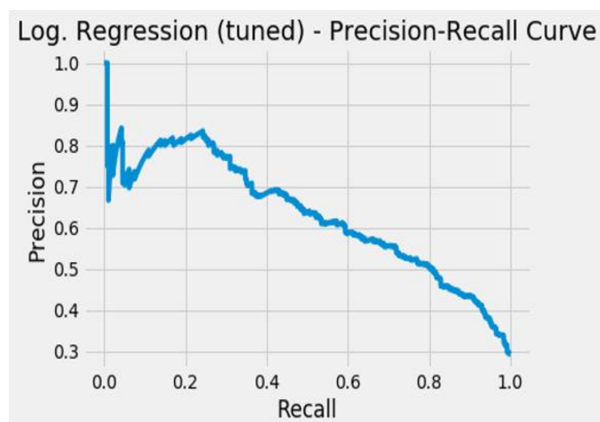
The datasets of customer are examined to comprise the equations of regression and an evaluation process for every customer in the set of data is then carried out. A consumer is at a risk of churn if the value of p for consumer is larger than a predefined value.






```
Accuracy Score Test: 0.7931769722814499
Accuracy Score Train: 0.8072888888888888 (as comparison)
```






```
AUC Score (ROC): 0.832353200014495
```

```
F1 Score: 0.5726872246696034
AUC Score (PR): 0.6317199512660256
```

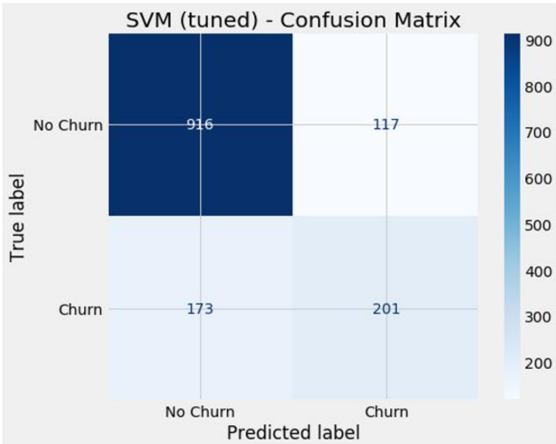
3.3 Random Forest

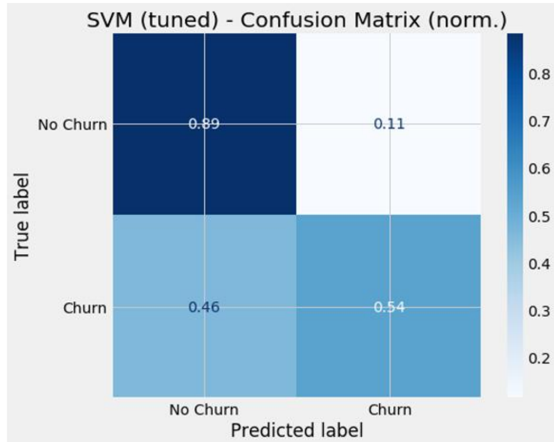
We applied Random Forest on the Training data set to validate if any further improvement of the model can be performed post the linear regression. Below were the parameters which were applied for Random Forest

For the Random Forest model Randomized Search CV is used to optimize for several hyper parameters including n_estimators, max_features, max_depth, criterion and bootstrap .Random Forest algorithm was also trained, we optimized the number of trees hyper parameter. We experimented with building the model by changing the values of this parameter every time in 100, 200, 300, 400 and 500 trees. Te best results show that the best number of trees was 200 trees. Increasing the number of trees after 200 will not give a signifcant increase in the performance. GBM algorithm was trained and tested on the same data, we optimized the number of trees hyper-parameter with values up to 500 trees. Te best value after the experiment was also 200 trees. GBM gave better results than RF and DT.

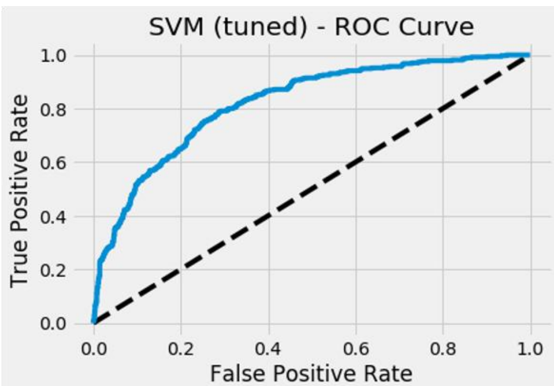
3.4 Support Vector Machine

Support Vector Machine (SVM) is a supervised ML algorithm which solves both regression and classification problems . The algorithm plots each data point in an n-dimensional space (n represents the number of features) where the value of each feature corresponds to the value of each coordinate. Data classification involves finding the most optimal hyperplane differentiating the classes perfectly.Effective in high dimensional,still effective in cases where number of dimensions is greater than the number of samples, uses a subset of training points in the decision function so it is also memory efficient, different kernel functions can be specified for the decision function.common kernels are provided,but it is also possible to specify custom kernals.

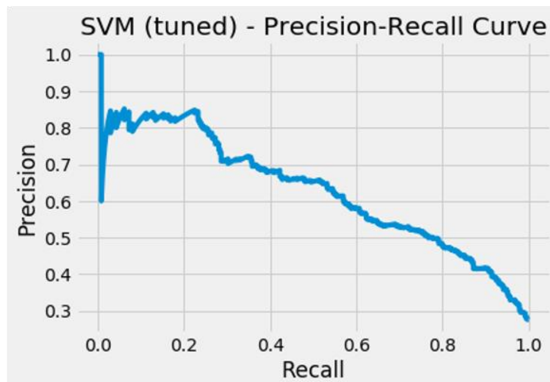





```
Accuracy Score Test: 0.7938877043354655
Accuracy Score Train: 0.8026666666666666 (as comparison)
```




```
AUC Score (ROC): 0.820055805478048
```




```
F1 Score: 0.5809248554913296
AUC Score (PR): 0.6282436588754954
```

3.5 Feed Forward Neural Network

Although the data set is relatively small and neural networks generally require lots of training data to develop meaningful prediction capabilities, a simple neural network is employed for a quick comparision to the other approaches.To address a potential bias stemming from the specific split of the data in the train-test-split part, cross-validation is used during hyper parameter tuning with Grid Search and Randomized Search. Cross validations splits the training data into in a specified amount of folds. For each iteration one fold is held out as “training dev” set and the other folds are used as training set. Result of cross-validation is k values for all metrics on the k-fold CV.

CHAPTER 4: KEY FINDINGS

Below table provides a snapshot of the various models which can be choose from based on the pros and cons of each model.

Sr.No	Model Name	Train (Accuracy Score)	Test (Accuracy Score)	AUC Score(ROC)	AUC Score(PR)	F1 Score
1	KNN	0.84 Optimized (0.80)	0.75 Optimized (0.76)	0.76 Optimized (0.79)	0.53 Optimized (0.53)	0.52 Optimized (0.53)
2	Logistic Regression	0.80 Optimized (0.80)	0.79 Optimized (0.79)	0.83 Optimized (0.83)	0.63 Optimized (0.63)	0.56 Optimized (0.57)
3	Random Forest	0.99 Optimized (0..88)	0.77 Optimized (0.79)	0.81 Optimized (0.82)	0.60 Optimized (0.64)	0.53 Optimized (0.55)
4	Support Vector Machine	0.81 Optimized (0.80)	0.78 Optimized (0.79)	0.77 Optimized (0.82)	0.58 Optimized (0.62)	0.55 Optimized (0.58)

The approach that finds the churn of customer are based on knowledge considering the company's calls and their clients. That data is stored in a table of database and is known as dataset. The pre-processing steps used in this study for dataset are: 1) first the spaces arereplacedwith values of null in the column of total charges; 2) the values of null are reducedfrom the column of total charges which comprises 15 percent missing data; 3) then thedata is converted to the type of float; 4) after than no internet service is replaced to no for the following columns: DeviceProtection, StreamingTV, OnlineSecurity, TechSupport, StreamingMovies and OnlineBackup; 5) the values for SenioCitizen is replaced with 0 as No and 1 as Yes; 6) Then the categorical column is made into Tenure; 7) After than the churn and nonchurn customers are separated; 8) Finally the numerical and categorical columns are separated. There are several measures available which can be used to verify the classification performance. The measures used in this study to verify the performance of classification are accuracy, recall, F-measure and precision. The significance of accuracy, recall, precision and F-measure is used to compare various classifiers effectiveness for prediction of churn. These metrics are applicable for examining any model performance which is constructed using both unbalanced and balanced set of data. Accuracy is defined as the accuracy prediction ratio to total set of predictions in a model. Precision is referred as an exactness measure.

These metrics are applicable for examining any model performance which is constructed using both unbalanced and balanced set of data. Accuracy is defined as the accuracy prediction ratio to total set of predictions in a model. Precision is referred as an exactness measure. It can also be referred as from the samples mentioned as positive and how many actually belongs to the positive set of attributes. Recall is regarded as completeness measure and it mentions about how many positive sample classes are classified properly.

CHAPTER 5: RECOMMENDATIONS AND CONCLUSION

5.1 Introduction

This chapter provides conclusion to the research topic "customer churn prediction in telco using in big machine learning data platform" followed by recommendations and suggestions based on the results of the research.

5.2 Conclusion

In the competitive telecom sector standardization and public policies of mobile communication permits customers to switch over from one carrier to another carrier easily resulting in a competitive market. The prediction of churn or the task of recognizing customers who are probable to discontinue service use is a lucrative and essential issue of telecom sector. Customer churn is often a critical problem for the telecom sector as customers do not delay to leave if they do not predict what they are viewing for. Customers mainly need value for money, competitive cost and greater service quality. Customer churning is associated directly to satisfaction of customer. It is a known fact that the customer acquisition cost is larger than customer retention cost that makes the retention a difficult prototype of business. There is no standard approach which resolves the churning problems of worldwide service providers of telco industry accurately. Data analytics with machine learning technique is used for customer churn which sets warning bells for customers before any damage could occur, providing telco firms the chance to take precautionary steps. These techniques are used to find the churn in customers by constructing models and studying from historical information. Conducting trials with perspective of end users, collecting their views on network, normalization of data, data set pre-processing, using feature selection, removing missing values and class imbalance and changing existing variables with derived variables develops the churn prediction accuracy which supports the telco sector to retain their customers much efficiently. It can be concluded that Big data analytics with machine learning were predicted to be an effective way for recognizing churn in customers.

It can be concluded that telecom service providers must provide much attraction to the above stated factors for their customers to reduce the churn rate flexibly and effectively. The cost of obtaining new customers can be greater than that of customer retention. One of the best way for customer retention is to reduce customers churn rate where churn refers to migration of customer from one service provider to another service provider or terminating particular services over particular periods for several reasons that can be predicted previously if the firm examines its records of data and uses machine learning technique which enhances the firms to find customers who are probable to churn. Several algorithms are available to reduce the churn rate in telecom companies. The telecommunication service providers use advanced analytics algorithms to mine through huge number of data of customers. This algorithm is smart enough to recognize hidden characteristics to find which customers are much probable to churn. Data mining plays an essential role in telecom firms and their effort to reduce overall churn develops good marketing strategies, recognize fraudulent activities and consumers and manage their network better. A proper algorithm is chosen relying on the problem nature and that of feasible data. It can be concluded that machine learning algorithms is regarded as one of the best solutions for telecom sector to reduce the churn rate.

5.3 Recommendations for future

It is recommended to expect behavioural patterns and customer churn. Telecom service providers must spend in insight tools and powerful analytics to expect churn of customers, finds behaviour of customer and devise strategies that enhances profitability as well as retention. The customer retention strategies costs must be mapped with expected return on interest to prioritize investments effectively. Telecom firms have to realign their priorities around retention of customers. It is recommended to employ co-browsing to provide a personalized service to customers. Be in person or on phone, telecom service providers must engage with a strong welcome message to customers which makes them feel appreciated and comfortable. Co-Browsing is one of the essential ways to add a personal feeling to consumer service. Quality service to customers is useful in reducing the churn rate of customers saving their effort in convincing customers to remain when they need to cancel. Co-browsing brings the customer representative and customer together on similar page offering a visual link and helping to build trust rapidly. It is recommended that telecom service providers must increase engagement of customers. In this competitive world customers are bombarded constantly by information and choices from all around. With the appropriate strategy of marketing in place and by concentrating on customer retention and satisfaction service providers must increase engagement of customers and nurture big term relations. Telecom service providers must implement tailored programs specifically to support their customers perceive the advantages of their services and products.

It is recommended that telecom service providers must delight and surprise their customers. A satisfied customer is the best strategy among all solutions to reduce the churn rate. Putting a smile on the face of customer is as easy as providing the best recognition award to customer. Telecom service provider must do something outstanding to show how much they value them.

Thus, the survival of any business is based on its capability to retain customers and put huge amount of efforts in reducing the churn rate of customers.

5.4 Summary

Customer churn is one of the major problems which the telecom sector is facing nowadays. It is essential to recognize possible customer churn so that the losses can be avoided. In order to maintain a loyal base of customer the service providers in telecom sector aims to retain customers with themselves. Since the costs related with obtaining a new customer is much greater than retaining older customer the prediction of churn becomes even more essential. The big data analysis with machine learning makes the churn prediction much easier in telecom sector. Thus, it can be concluded that the big data analytics with machine learning techniques have proven to be accurate and effective to predicts customer churn in nearby future.

CHAPTER 6: REFERENCES

[1] K-Nearest Neighbor

https://www.geeksforgeeks.org/k-nearest-neighbours/

[2] Feed Forward Neural Network

https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/

[3] One-Hot Encoding

https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

[4] Supervised Machine Learning Algorithms

https://www.javatpoint.com/supervised-machine-learning