International Journal of Computer Applications Technology and Research
Volume 8–Issue 09, 363-366, 2019, ISSN:-2319–8656

Customer Churn Analysis and Prediction

Aditya Kulkarni [1]
M.sc (Big Data Analytics)
MIT WPU
Pune , India

Amruta Patil [2]
Msc (Big Data Analytics)
MIT WPU
Pune , India

Madhushree Patil [3]
Msc (Big Data Analytics)
MIT WPU
Pune , India

Sachin Bhoite [4]
Assistant Professor , Computer Science
MIT WPU
Pune , India

Abstract: When talking about any companies growth within market customers play an essential role in it , having the correct insights about customer behaviour and their requirements is the current need in this customer driven market . Preserving the interests of customers by providing new services & products helps in maintaining business relations . Customer churn is great problem faced by companies nowadays due to lagging in understanding their behaviour & finding solutions for it . In this project we have found causes of the churn for a telecom industry by taking into consideration their past records & then recommending them new services to retain the customers & also avoid churns in future . We used pie charts to check churning percentage later analysed whether there are ant outliers [using box plot] then dropped some features which were of less importance then converted all categorical data into numerical by using [Label Encoding for multiple category data & map function for two category data] plotted the ROC curve to get to know about true positive & false negative rate getting line at 0.8 then spitted the data using train test split .We used algorithms decision tree , Random Forest for feature selection wherein we got feature importance , then used logistic regression & found feature with highest weight assigned leading to cause of churn . Now in order to retain customers we can recommend them new services.

Keywords : Customer churn analysis telecom , Customer churn prediction & prevention , naïve bayes , logistic regression , decision tree , random forest

1.INTRODUCTION

The telecom industry is growing day by day hence user as well as operators are investing into this industry ,such a customer driven industry faces a huge financial issue if customer tend to leave their services . By using machine learning we can analyse , predict the way customer respond to these services , researches have proven that by using past data it could be accomplished [2] .

In this Customer Churn prediction & retention we are analysing the past behaviour of customers and accordingly finding the real cause of the churn , then predicting whether churn will happen in future by customers . By taking into account details like Monthly charges , services they have subscribed for , tenures , contract they will contribute into he end results i.e prediction.

Our aim is to use machine learning concepts to not only predict & retain customers but also to avoid further churns which would be beneficial to industry .

2.RELATED WORKS

We went through various articles & research papers , and then found that many researchers have worked on customer churn as it is a major problem faced by industries nowadays we found the following papers more promising

"A comparison of machine learning techniques for customer churn prediction Praveen Asthana has used decision tree , svm , naïve bayes , ANN & compared which model gives best accuracy and would help in prediction of customer churn to achieve better performance[1].

SCOTT A. NESLIN, SUNIL GUPTA, WAGNER KAMAKURA, JUNXIANG LU, and CHARLOTTE H. MASON* "Defection Detection: Measuring and

Understanding the Predictive Accuracy of Customer Churn Models" [2]here they have worked on measuring and increasing accuracy for churn prediction used logistic & tree approach .

We went through one more paper "Customer churn prediction in telecom using machine learning in big data platform" Abdelrahim Kasem Ahmad* , Assef Jafar and Kadan Aljoumaa [3] they have used decision tree , random forest , XGBoosting , they used this algorithm for classification in predictive churn of customers getting better accuracy.

S-Y. Hung, D. C. Yen, and H.-Y. Wang. "Applying data mining to telecom churn management." , here they have used predictive model in a bank with personalized action to retain customer & have also used recomender system[4] .

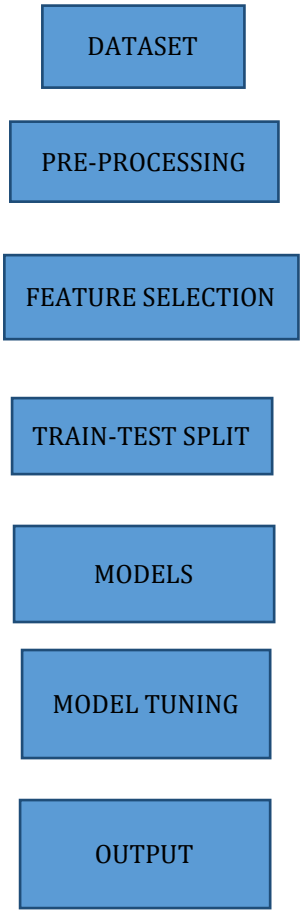
K. Coussement, and D. Van den Poel "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers.", they have used logistic regression , svm & random forest classification algorithms to filter out the churners from non-churner[5].

L Miguel APM. "Measuring the impact of data mining on churn management" , they have proposed a analysis framework which prefigure impact of data mining for churn management[6] .

Adnan Amin , Babar shah , Awais Adnan "Customer churn prediction in telecommunication industry using data certainty"[7], The dataset is grouped into different zones based on the distance factor which are then divided into two categories as data with high certainty, and data with low certainty, for predicting customers exhibiting Churn and Non-churn behaviour.

3. PROCESS FLOW

The data we got was mostly balanced & categorical data then we began with Data Cleaning, Pre-processing, removing unwanted columns, feature selection, label encoding.



3.1 DATASET

We took this telecom dataset from online website source took all the insights regarding the data .
Attributes of the dataset : :
Customerid, gender, SeniorCitizen, Partner, Dependents,tenure, PhoneService, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, Churn.

3.2 DATA PRE-PROCESSING

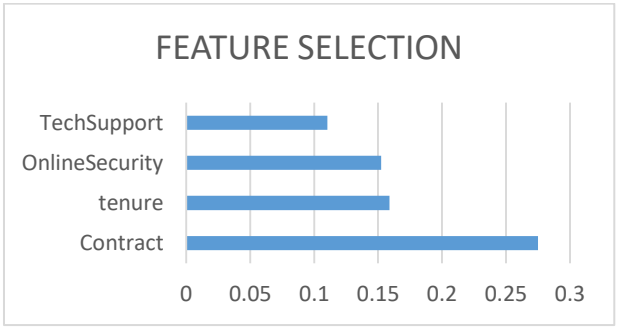
Data pre-processing is important task in machine learning. It converts raw data into clean data. Following are technique, we have applied on data: -

- Missing Values – Here we had missing values in Totalcharges feature which we then eliminated and adjusted them with mean values . These are the missing row values within data if not handled would later lead to errors for converting data type as it takes string value for empty spaces .
- Label Encoder – For categorical variables this is perfect method to convert them into numeric values , best used when having multiple categories . We had various categorical values converted them into numeric for further use in algorithms .
- Drop Columns – As we took insights from the data we came to know some of the features were of less importance so we dropped them to reduce number of features .

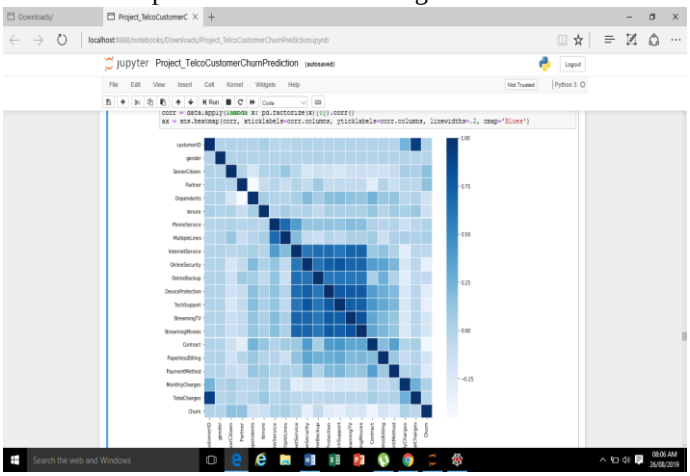
3.3 FEATURE SELECTION

As we had number of features and most of them were of great importance so we used feature section to get to know which of them are contributing towards the accuracy of the model .

We used Decision tree , Random forest for feature selection so using decision tree we got accuracy[80] and by using arandom forest we go t [80%] so random forest gave us four features
Index(['tenure', 'Contract', 'MonthlyCharges', 'TotalCharges'], dtype='object')
[(0.2251735641431145, 'Contract'), (0.1687558104226648, 'tenure'), (0.12539865168020692, 'OnlineSecurity'), (0.1128092761196452, 'TechSupport'), (0.10731999001345587, 'TotalCharges'), (0.08573112448285626, 'MonthlyCharges'),
Here we can see contarct is having more importance resulting factor for churn .

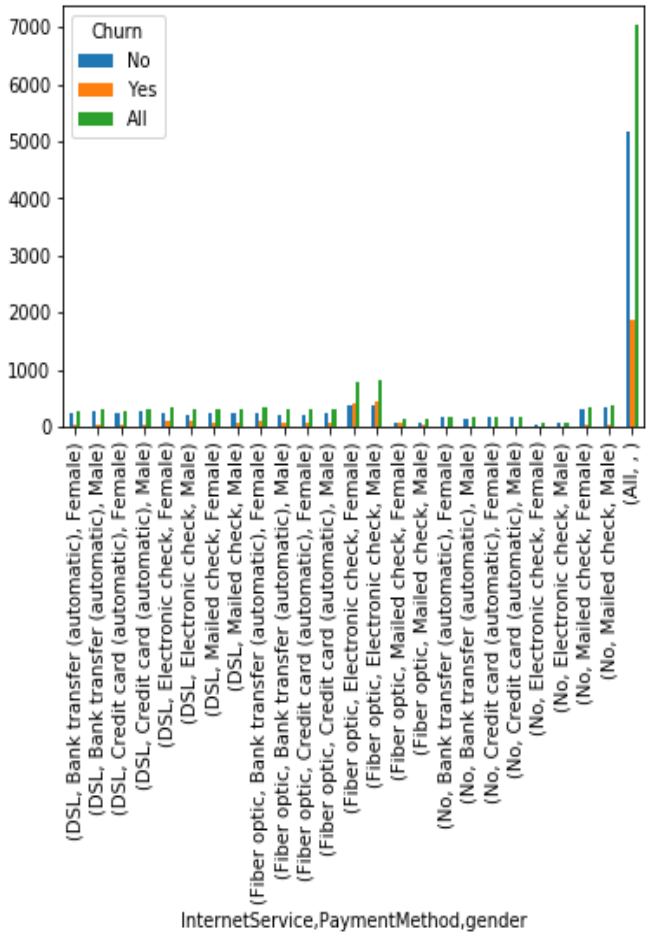


And for decision tree we got accuracy of [77%]
Used heat Map for correlation checking :

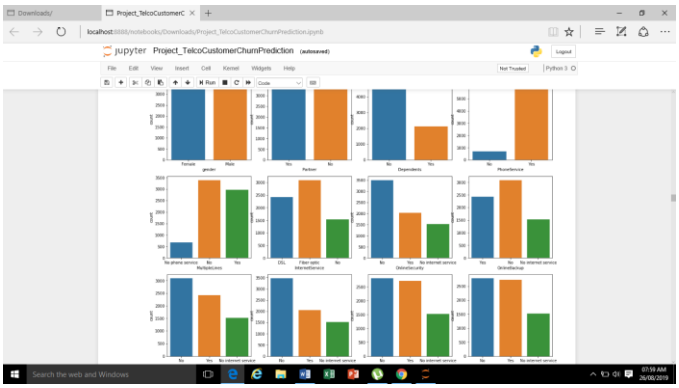


4. EXPLORATORY DATA ANALYSIS

In this phase we will look towards those features which we didn't consider in feature selection but are contributing factor for prediction .

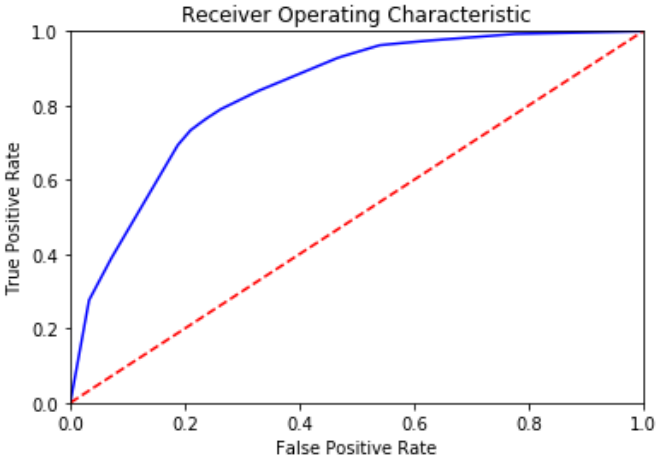


Here we can see that customer who took fibre optics for month-to-month contract whether it be male/female resulted in churn.



Also visualised all the features within the dataset & came to know the distributions .

Got roc curve ;



Confusion Matrix :
[[3816 295]
[924 590]] .

5. RESULT AND DISCUSSION

Now after all the cleaning up & pre-processing of the data now we separate our data for further applying algorithms on it. By using :

1. Train-Test Split
2. Modeling
3. Tuning Model

5.1 Train-Test Split:

To create the model we train our dataset while testing data set is used to test the performance. So, in our data, we have split into 80% for training data and 20% for testing data because it makes the classification model better whilst more test data makes the error estimate more accurate.

5.2 Modelling :

Following are model, we applied to check which model gives better accuracy:

- Support Vector Classifier (SVC):
  This algorithm is used for classification problem. The main objective of SVC is to fit to the data you provide, returning a "best fit" hyperplane that divides, or categorizes, your data. From there, when obtaining the hyperplane, you'll then feed some options to your category to examine what the "predicted" class is.

- Decision Tree:
  Decision tree is non-parametric supervised learning. It is used for both classification and regression problem. It is flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The path between root and leaf represent classification rules. It creates a comprehensive analysis along with each branch and identifies decision nodes that need further analysis.

- Random Forest:
  Random Forest is a meta estimator that uses the number of decision tree to fit the various sub samples drawn from the original dataset. we also can draw data with replacement as per the requirements.

- K-Nearest Neighbours (KNN):

K-Nearest Neighbours (KNN) is supervised learning algorithm which is used to solve regression and classification problem both. Where 'K' is number of nearest neighbours. It is simple to implement, easy to understand and it is lazy algorithm. Lazy algorithm means it does not need any training data points for model generation . All training data used in the testing phase.

- Naïve Bayes:
  A Naive Bayes Classifier is a supervised machine learning algorithm which uses the Bayes' Theorem, that features are statistically independent. It finds many uses in the probability theory and statistics. By simple machine learning problem, where we need to learn our model from a given set of attributes (in training examples) and then form a hypothesis or a relation to a response variable.

  - Logistic Regression :

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes . Logistic Regression is a Machine Learning algorithm which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability.

Models used & their accuracy : :

Model	Accuracy
Logistic Regression	80.38%
Decision Tree	77.81%
Random Forest Tree	80.02%
Naïve Bayes	74.91%
SVM	80.1%
K – Nearest Neighbour	76.61%
XGBoost	80%

Figure 5: Accuracy for Different Models

5.3 MODEL TUNING:

Here we tune the model to increase model performance without overfitting the model.

- XGBoost **:**
  XGBoost stands for extreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance[3].

We used XGBoost to check the error function and reduce it [Accuracy :80%] , by using cross validation checked for the reducing RMSE also it handles the missing values , initially our RMSE was [[0]validation_0-error: 0.208955] later it came [[10] validation_0-error: 0.200426]

6. CONCLUSION

Here we had past records of customers who had churned and using that data we predicted whether new customer would tend to churn or not , this will help the companies to get to know the behaviour of customer & how to maintain their interests into the services of company . Further the company can also use recommender system to retain customers and also avoid the further churns . We used various algorithms wherein Logistic regression gave us high accuracy close to this accuracy were Random Forest , SVM .

The dataset did not consisted of records which would tell us whether customer has switched the services , that will help in recommending new services further . Now we are going to build a recommender system to avoid churns & retain the old customers .

7. REFERENCES

[1] Praveen Ashtana "A comparison of machine learning techniques for customer churn prediction" International Journal of Pure and Applied Mathematics Volume 119 No. 10 2018, 1149-1169 ISSN: 1311-8080

[2] SCOTT A. NESLIN, SUNIL GUPTA, WAGNER KAMAKURA, JUNXIANG LU, and CHARLOTTE H. MASON* "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models" Journal of Marketing Research 204 Vol. XLIII (May 2006), 204–211 , ISSN: 0022-2437.

[3] Abdelrahim Kasem Ahmad* , Assef Jafar and Kadan Aljoumaa "Customer churn prediction in telecom using machine learning in big data platform" - Journal of Big Data **volume 6**, Article number: 28 (2019) , published on 20th March 2019 .

[4] S-Y. Hung, D. C. Yen, and H.-Y. Wang. "Applying data mining to telecom churn management." Expert Systems with Applications, vol. 31, no. 3, pp. 515–524, 2006.

[5] K. Coussement, and D. Van den Poel. "Improving customer attrition prediction by integrating emotions from client/company interaction emails and evaluating multiple classifiers." Expert Systems with Applications, vol. 36, no. 3, pp. 6127–6134, 2009

[6] L. Miguel APM. "Measuring the impact of data mining on churn management." Internet Research, vol. 11, no. 5, pp. 375–387,2001

[7] Amin , Babar shah , Awais Adnan "Customer churn prediction in telecommunication industry using data certainty" Journal of business research Volume 94, January 2019, Pages 290-301.