International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Customer Attrition Classification and EDA on IBM Telecommunication Dataset based on Machine Learning Algorithms

S Shashank Raj[1], Akhil Chauhan [2], B Madhu[3], Vairachilai S [4]

[1,2]Student, Department of Computer Science,
[3,4]Professor, Department of Computer Science and Engineering, Faculty of Science and Technology (IcfaiTech), The ICFAI Foundation for Higher Education (IFHE), Telangana, India

---

Abstract - *Customer churn or attrition is a crucial problem and one among the foremost principal concerns for giant telecom companies. Thanks to the undeviating effect on the revenues of the telecom companies, they're seeking to evolve the means to predict a customer to churn. Therefore, finding factors that increase customer attrition may be a major issue to require necessary actions to scale back this attrition. The principal contribution of our work is to develop a customer attrition prediction model which helps telecom operators to predict customers who are subjected to churn from that operator. The model developed during this work uses machine learning classification algorithms. We use the data provided by IBM Telecom Company from Kaggle. The dataset contained all customers' information over the company, and was split to train and test. The developed model experimented six algorithms: Logistic Regression, Decision Tree, Random Forest, Ada Boost, XG Boost, K Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machine(SVM). All the developed models later used for prediction using test data and some metrics has been taken to determine the optimum attrition model.*

Key Words: Machine Learning, Customer Attrition, Logistic Regression, Decision Tree, SVM.

1. INTRODUCTION

Telephone operator companies, Internet service providers, TV broadcasting channel companies, often use customer attrition or churn analysis and customer attrition rates are one among the major business metrics because the value of retaining an existing customer is way but acquiring a replacement one. Companies from these sectors often have customer service branches which strive to get back turn traitor clients, because recuperated long-term customers are often worth far more to a corporation than newly joined clients.

Telecom Companies usually make a difference between voluntary attrition and involuntary attrition. Voluntary attrition occurs due to the choice by the customer to change to a different operator or service provider, involuntary attrition occurs due to the circumstances like a customer's relocation to a long-term care facility, death, or the relocation to a different country. In major applications, involuntary reasons of customers for churn from operator are excluded from the analytical models. Analysts tend to consider voluntary attrition, because it typically occurs due to factors of the company-customer relationship which companies control over the period, like how the network connections are or how billing interactions are handled or how after-sales assistance is provided.

We focused on evaluating and analyzing the performance of the applied machine learning algorithms for predicting churn in telecommunications companies. We experimented a number of algorithms such as Logistic Regression, Decision Tree, Random Forest, Ada Boost, XG Boost Naïve Bayes, Support Vector Machine and K Nearest Neighbors to build the predictive model of customer Attrition after data preparation, feature engineering, and feature selection methods. Many researchers confirmed that machine learning technology is highly efficient to predict this situation, thus Machine Learning algorithms are applied through learning from previous data; The data used contains 7043 rows (customers) and 21 columns (features or attributes).

2. LITERATURE SURVEY

Numerous approaches have been applied to predict churn in telecom companies. Most of those approaches have used machine learning and data processing. The major part of the related work is focused on applying just one method of knowledge mining to extract patterns, and therefore the others focused on comparing several strategies to predict attrition.

Gavril et al. [5] presented a complicated methodology of knowledge mining to predict churn for prepaid customers using dataset for call details of 3333 customers with 21 features, and a dependent churn parameter with two values: Yes/No. Some features include information about the amount of incoming and outgoing messages and voicemail for every customer.

The author applied principal component analysis algorithm "PCA" to scale back data dimensions. Three machine learning algorithms were used: Neural Networks, Support Vector Machine, and Bayes Networks to predict churn factor. The author used AUC to live the performance of the algorithms. The AUC values were 99.10%, 99.55% and 99.70% for Bayes Networks, Neural networks and support vector machine, respectively. The dataset utilized in this study is little and no missing values existed.

He et al. [6] proposed a model for prediction supported the Neural Network algorithm so as to unravel the matter of customer churn during a large Chinese telecom company which contains about 5.23 million customers. The prediction accuracy standard was the general accuracy rate, and reached 91.1%.

Makhtar et al. [7] proposed a model for churn prediction using rough pure mathematics in telecom. As mentioned during this paper Rough Set classification algorithm outperformed the opposite algorithms like rectilinear regression, Decision Tree, and Voted Perception Neural Network.

We didn't find any research curious about this problem recorded in any telecommunication company. Majority of the previous research work didn't perform the feature engineering phase or build features from data while they depended on the ready attributes provided either by telecom companies or published online. During this paper, the feature engineering phase is taken into consideration to make our own features to be utilized in machine learning algorithms.

3. DESCRIPTION OF DATA

Attributes	Description	Type of Data
Customer ID	Customer Identification Number	Numerical
Gender	Whether the customer is a male or a female	Categorical
Senior Citizen	Whether the customer is a senior citizen or not	Yes, No
Partner	Whether the customer has a partner or not	Yes, No
Dependents	Whether the customer has dependents or not	Yes, No
Tenure	No of months the customer has stayed with the company	Numerical
Phone Service	Whether the customer has a phone service or not	Yes, No
Multiple Lines	Whether the customer has multiple lines or not	Yes, No, NPC
Internet Service	Customer's internet service provider	DSL, Fiber optic, No
Online Security	Whether the customer has online security or not	Yes, No, NIS
Online Backup	Whether the customer has online backup or not	Yes, No, NIS
Device Protection	Whether the customer has device protection or not	Yes, No, NIS
Tech Support	Whether the customer has tech support or not	Yes, No, NIS
Streaming TV	Whether the customer has streaming TV or not	Yes, No, NIS
Streaming Movies	Whether the customer has streaming movies or not	Yes, No, NIS
Contract	The contract term of the customer	Per Month, 1 or 2 year
Paperless Billing	Whether the customer has paperless billing or not	Yes, No
Payment Method	The customer's payment method	EC, MC, BT, CC
Monthly Charges	The amount charged to the customer monthly	Numerical
Total Charges	The total amount charged to the customer	Numerical
Churn	Whether the customer churned or not	Yes, No

Table -1: Description of Attributes
*NPS - No Phone Service, NIS - No Internet Service, EC - Electronic Check, MC - Mailed check,
BT - Bank transfer, CC - Credit card

4. EXPLORATORY DATA ANALYSIS (EDA)

4.1 Customer Churn Rate

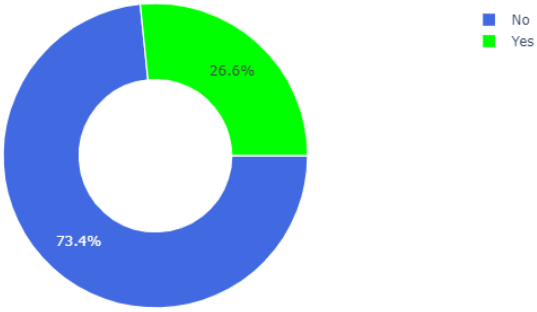

Chart -1: The Customer Churn Percentage

About a quarter of the Customers have been leaving this telecommunication company, which is very bad rate. This includes Voluntary and Involuntary Churn

4.2 Tenure Group Distribution in Customer Attrition

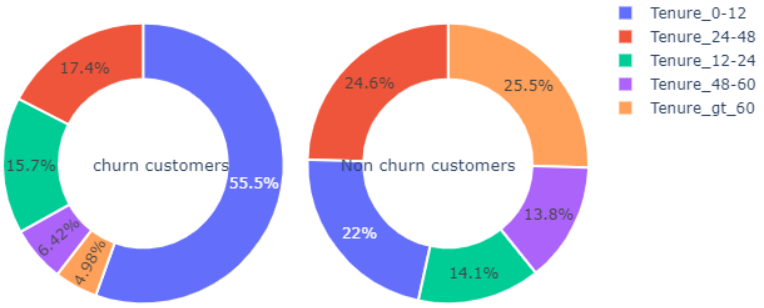

Chart -2: Tenure Group Distribution

About half of the Customers who are leaving the company are the customers who stayed about a year; the customers who joined about 60 months ago are only 4.98%

4.3 Total Charges Distribution in Customer Attrition

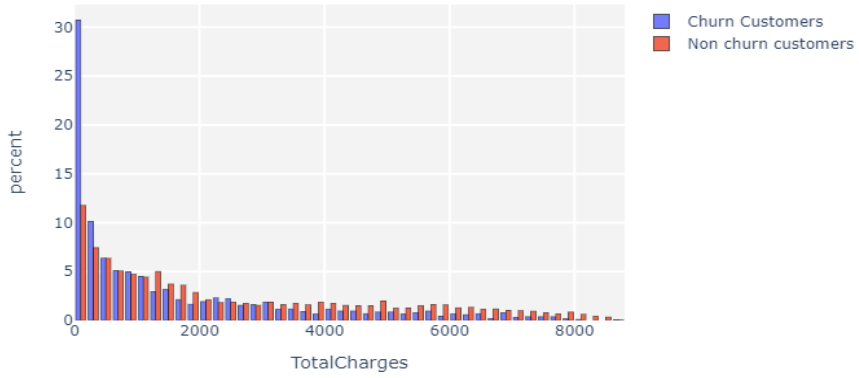

Chart -3: Tenure Group Distribution

Most of the Customers who are leaving the company have zero total charges

4.4 Payment Method Distribution in Customer Attrition

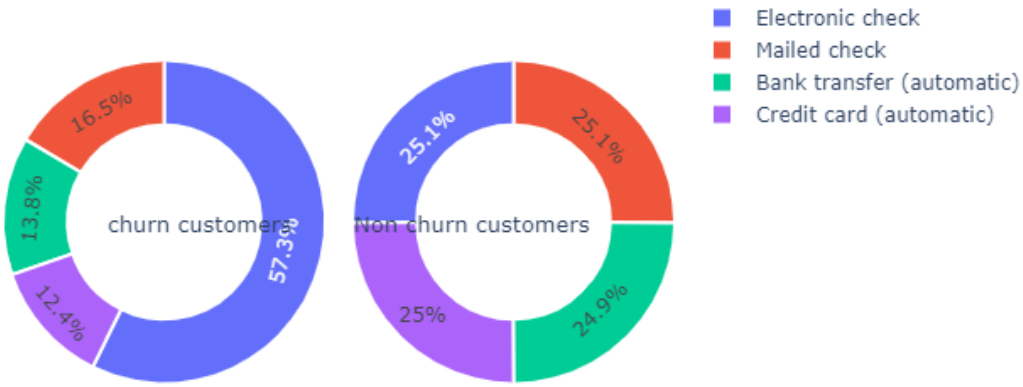

Chart -4: Payment Method Distribution

Most of the Customers who are leaving are paying their bills through Electronic Check

4.5 Paperless Billing Distribution in Customer Churn

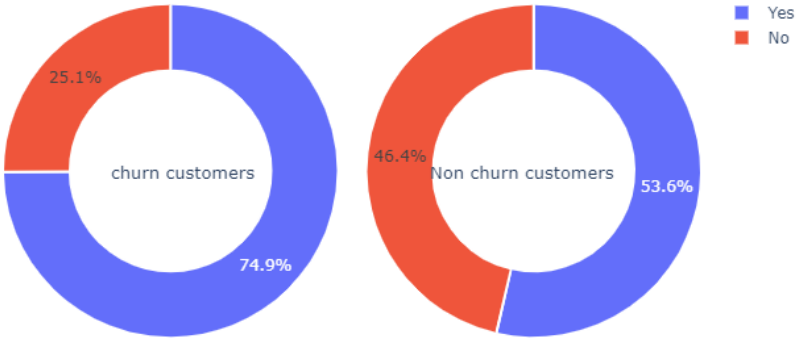

Chart -5: Paperless Billing Distribution

Most of the Customers who are leaving are paying their bills Online

4.6 Scatter Plot Matrix for Numerical Columns for Customer Attrition

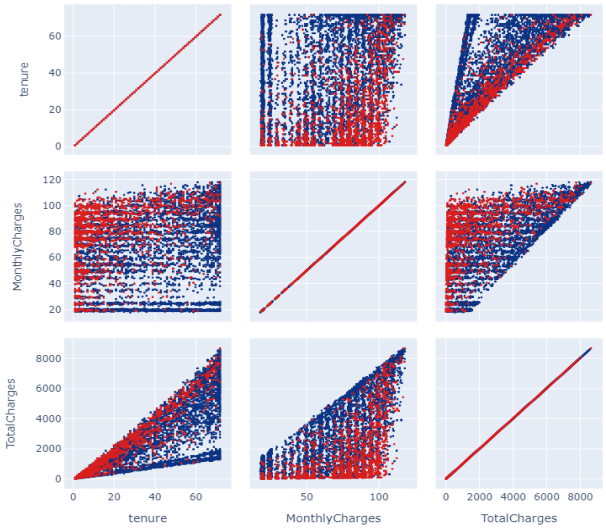

Chart -6: Scatter Plot for all Numerical Attributes

4.7 Contract Distribution in Customer Attrition

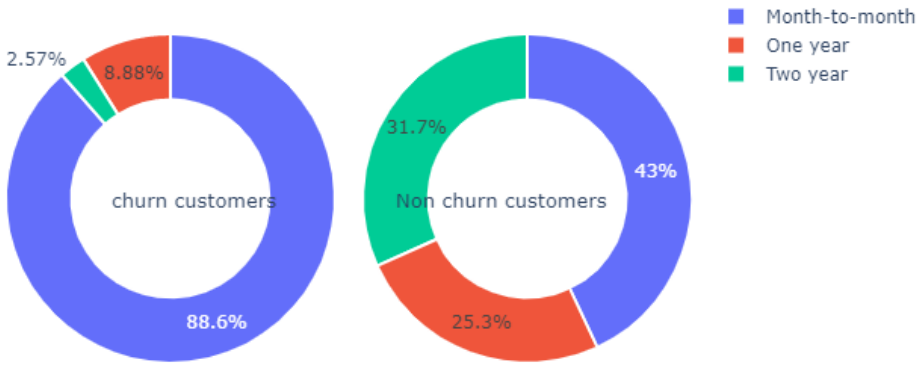

Chart -7: Contract Distribution

Most of the Customers who are leaving the company are on monthly contract basis

4.8 Tech Support Distribution in Customer Attrition

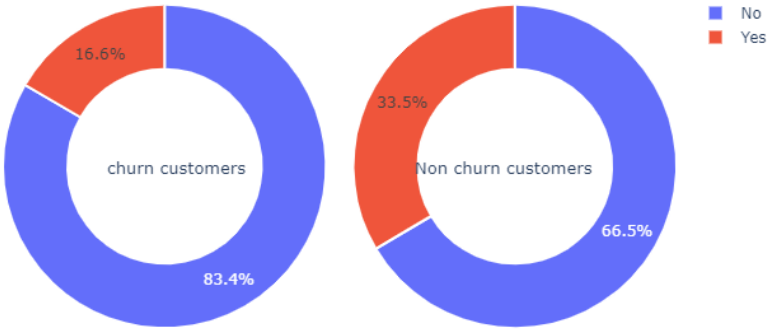

Chart -8: Tech Support Distribution

Most of the Customers who are leaving the company are not on Tech Support

4.9 Correlation Matrix

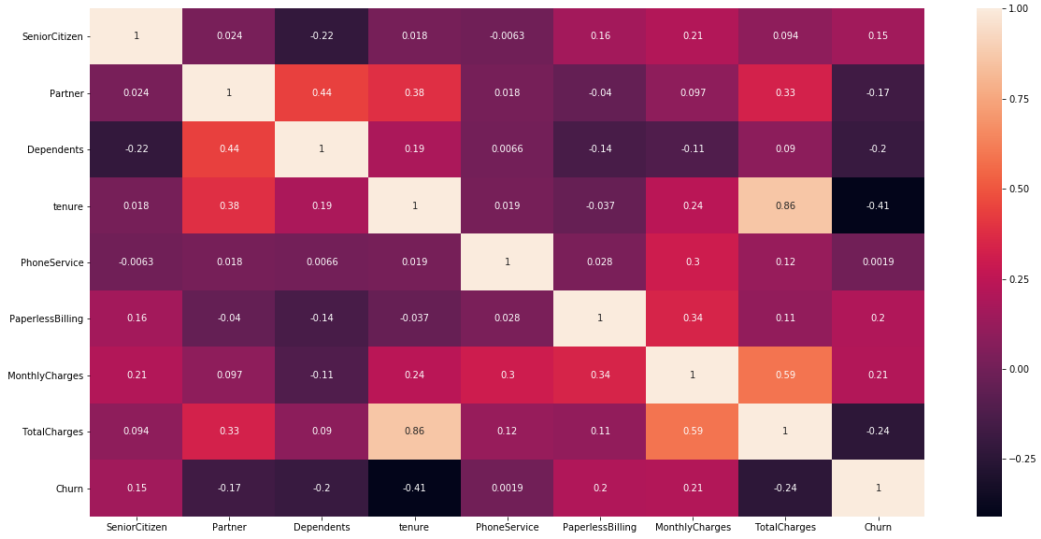

Fig -1 : Correlation Matrix for all the attributes

Irjet Template sample paragraph .Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

5. EVALUATION METRICS

5.1 Confusion Matrix

Logistic Regression		
	Predicted Value	
Actual Value	Yes	No
Yes	1418	162
No	244	286

Table -2: Confusion Matrix of Logistic Regression

Random Forest		
	Predicted Value	
Actual Value	Yes	No
Yes	964	88
No	181	174

Table -3: Confusion Matrix of Random Forest

SVM		
	Predicted Value	
Actual Value	Yes	No
Yes	953	89
No	164	201

Table -4: Confusion Matrix of SVM

Ada Boost		
	Predicted Value	
Actual Value	Yes	No
Yes	958	84
No	175	190

Table -5: Confusion Matrix of Ada Boost

XG Boost		
	Predicted Value	
Actual Value	Yes	No
Yes	939	103
No	165	200

Table -6: Confusion Matrix of XG Boost

KNN		
	Predicted Value	
Actual Value	Yes	No
Yes	863	179
No	178	187

Table -7: Confusion Matrix of KNN Classifier

Naïve Bayes		
	Predicted Value	
Actual Value	Yes	No
Yes	675	367
No	58	307

Table -8: Confusion Matrix of Naïve Bayes Classifier

Decision Tree		
	Predicted Value	
Actual Value	Yes	No
Yes	859	183
No	177	188

Table -9: Confusion Matrix of Decision Tree

5.2 Metrics

Logistic Regression				
	Precision	Recall	F1 – Score	Support
0	0.85	0.90	0.87	1580
1	0.64	0.54	0.58	530
Accuracy			0.81	2110
Macro Avg	0.74	0.72	0.73	2110
Weighted Avg	0.80	0.81	0.80	2110

Table -10: Evaluation Metrics of Logistic Regression

Random Forest				
	Precision	Recall	F1 – Score	Support
0	0.84	0.92	0.88	1052
1	0.66	0.49	0.56	355
Accuracy			0.81	1407
Macro Avg	0.75	0.70	0.72	1407
Weighted Avg	0.80	0.81	0.80	1407

Table -11: Evaluation Metrics of Random Forest Classifier

SVM				
	Precision	Recall	F1 – Score	Support
0	0.85	0.91	0.88	1042
1	0.69	0.55	0.61	365
Accuracy			0.82	1407
Macro Avg	0.77	0.73	0.75	1407
Weighted Avg	0.81	0.82	0.81	1407

Table -12: Evaluation Metrics of SVM

Ada Boost				
	Precision	Recall	F1 – Score	Support
0	0.85	0.92	0.88	1042
1	0.69	0.52	0.59	365
Accuracy			0.82	1407
Macro Avg	0.77	0.72	0.74	1407
Weighted Avg	0.81	0.82	0.81	1407

Table -13: Evaluation Metrics of Ada Boost

XG Boost				
	Precision	Recall	F1 – Score	Support
0	0.85	0.90	0.88	1042

1	0.66	0.55	0.60	365
Accuracy			0.81	1407
Macro Avg	0.76	0.72	0.74	1407
Weighted Avg	0.80	0.81	0.80	1407

Table -14: Evaluation Metrics of XG Boost

KNN Classifier				
	Precision	**Recall**	**F1 – Score**	**Support**
0	0.83	0.83	0.83	1042
1	0.51	0.51	0.51	365
Accuracy			0.75	1407
Macro Avg	0.67	0.67	0.67	1407
Weighted Avg	0.75	0.75	0.75	1407

Table -15: Evaluation Metrics of KNN Classifier

Naïve Bayes				
	Precision	**Recall**	**F1 – Score**	**Support**
0	0.92	0.65	0.76	1042
1	0.46	0.84	0.59	365
Accuracy			0.70	1407
Macro Avg	0.69	0.74	0.68	1407
Weighted Avg	0.80	0.70	0.72	1407

Table -16: Evaluation Metrics of Naïve Bayes Classifier

Decision Tree				
	Precision	**Recall**	**F1 – Score**	**Support**
0	0.83	0.82	0.83	1042
1	0.51	0.52	0.51	365
Accuracy			0.74	1407
Macro Avg	0.67	0.67	0.67	1407
Weighted Avg	0.75	0.74	0.74	1407

Table -17: Evaluation Metrics of Decision Tree

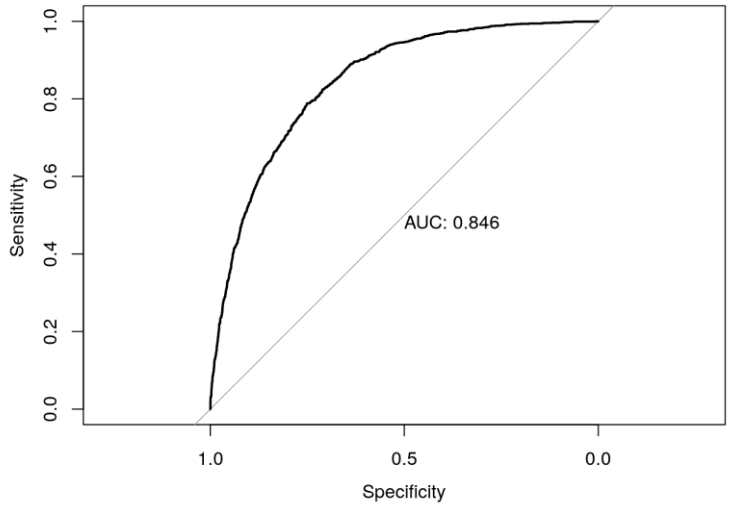

Fig -2: ROC Curve of Logistic Regression

6. CONCLUSIONS

Accuracies of Logistic Regression, Random Forest, SVM, Ada Boost, XG Boost, KNN, Naïve Bayes and Decision Tree are 80.75%, 80.88%, 82.01%, 81.59%, 80.95%, 74.62%, 69.79% and 74.41% respectively. It is important to scale the variables in logistic regression so that all of them are within a range of 0 to 1. This helped to improve the accuracy from 79.7% to 80.7%.

The logistic regression model and random forest model work better than the decision tree model. The Accuracies are 0.78 for Logistic Regression, 0.78 for Decision Tree and 0.79 for Random Forest, with 0.5 as threshold. From random forest algorithm, monthly contract, tenure and total charges are the most important predictor variables to predict churn. The results from random forest are very similar to that of the logistic regression and in line to what we had expected from our EDA. With SVM we were able to increase the accuracy to up to 82%. Interestingly with XG Boost we were able to increase the accuracy on test data to almost 82%. Clearly, XG Boost is a winner among all other techniques. But XG Boost is a slow learning model and is based on the concept of Boosting.

Logistic regression and SVM classification algorithms have the highest accuracy. But our data is imbalanced. So it is important to look at the confusion matrix according to these two algorithms. With imbalanced datasets, the highest accuracy does not give the best model. Assume we have 1000 total rows, 10 rows are churn and 990 rows are non-churn. If we find all these 10 churn rows as non-churn, then the accuracy will be still 99%. Although it is a wrong model, if we do not look at the confusion matrix, then we cannot identify the error.

Regarding the variance importance, the logistic regression model and the random forest model have little differences. They both have Monthly Charges, tenure, Contract and Payment Method as important predictors and have gender, Streaming TV, Movies and Partner as unimportant predictors. However, in the logistic regression model, Paperless Billing, Phone Service and Online Backup show significant influence on the churn, while in the random forest model, they have very small predicting power.

We can see that some variables have a negative relation to our predicted variable (Churn), while some have positive relation. Negative relation means that likeliness of churn decreases with that variable. As we saw in our EDA, having a 2 month contract reduces chances of churn. 2 month contract along with tenure have the most negative relation with Churn as predicted by logistic regressions, having DSL internet service also reduces the probability of Churn, lastly, total charges, monthly contracts, fiber optic internet services and seniority can lead to higher churn rates. This is interesting because although fiber optic services are faster, customers are likely to churn because of it. I think we need to explore more to better understand why this is happening.

Since data set is imbalanced, we preferred to use the F1 score rather than accuracy. Logistic Regression gives the highest F1 Score, so it is the best model, Naive Bayes is the worst model because it gives the lowest F1 score. Sex has no impact on churn. People having month-to-month contract tend to churn more than people having long term contracts. As the tenure increases, the probability of churn decreases. As monthly charges increases, the probability of churn increases.

7. REFERENCES

[1] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016.p. 97-100.

[2] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In : Sixth international conference on fuzzy systems and knowledge discovery, vol.1.2009.p.92-4

[3] Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory, J Fundam Appl Sci. 2017;9(6):854-68.

[4] Burez D, den Poel V, Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36(3):4626-36.

[5] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.

[6] He Y, He Z, Zhang D. A study on prediction of customer churn in fixed communication network based on data mining. In: Sixth international conference on fuzzy systems and knowledge discovery, vol. 1. 2009. p. 92–4.

[7] Makhtar M, Nafis S, Mohamed M, Awang M, Rahman M, Deris M. Churn classification model for local telecommunication company based on rough set theory. J Fundam Appl Sci. 2017;9(6):854–68.

[8] Chawla N. Data mining for imbalanced datasets: an overview. In: Data mining and knowledge discovery handbook. Berlin: vvSpringer; 2005. p. 853–67.

[9] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access. 2016;4:7940–57.

8. ACKNOWLEDGEMENT

Firstly we would like to thank Madhu Sir, for giving me the opportunity to complete our Special Project under his supervision, it is truly an honor. Thank you for all the advice, ideas, moral support and patience in guiding me through this project. I also want to thank my friends Vamshidhar Reddy, Kosuri Vijay Kumar and Murali Reddy for helping me to complete this project.

BIOGRAPHIES

S SHASHANK RAJ
Student at
ICFAI Tech School

AKHIL CHAUHAN
Student at
ICFAI Tech School

MADHU B
Professor at
ICFAI Tech School

S VAIRACHILAI
Professor at ICFAI Tech School