Customer Churn Prediction

Loyola Marymount University

Hetanshee Shah, Jeet Desai, Mustafa Zaki

***Abstract*— One of the most important parts of running any business is to understand the value of your customers. And in order to survive, or even thrive in your business is to identify customers who are not hesitant to leave your business and turn towards your competitor. Customer churn model aims to identify this and generate a binary value - indicating whether a customer will churn or not. In this paper customer information from an Iranian Telecommunication company was used to predict if a customer will churn or not. The key steps used to implement this is: data preprocessing - where we clean the data from various impurities, model construction - building 3 models that will be used to predict if a customer will churn or not, and lastly, analysis of the data was done in order to mitigate the customer churn rate.**

I. INTRODUCTION

The mobile services market is growing significantly and sustainably, not only due to the size of the market, but also due to the increasing variety of services offered and fierce competition in the telecommunication industry. Regardless of the earliest stages of this industry, the method of contest has moved from procuring new endorsers to holding existing customers. This has been accomplished by participating in showcasing endeavours and by luring customers from rival organizations.

Based on a Jan 2020 article, Accenture reports that 77% of consumers now retract their loyalty more quickly than they did three years ago and industry must therefore work harder than ever to retain their customer bases. Acquisition costs far outweigh those of keeping current customers, further motivating companies to implement innovative strategies to boost customer retention in the telecom industry. This is further underscored by research from Bain Company suggesting that a mere 5% increase in a company's retention rate can increase profits by 25% to 95%. Hence, financially, it makes more sense for an organization to focus on retaining its existing customers. As a result, churn management is a major area of focus.

In this study to predict the customer churn rate, the imbalanced data is treated with different classifiers such as decision tree, logistic regression, and support vector classifier. The results are then compared based on different performance metrics such as Precision, Recall, F1 score and auc value.

A. DATASET

The data set has been collected from an Iranian telecommunication company's database over a period of 12 months. It contains 3150 customer data with the following features:

Feature Name	Type	Description
Call Failures	Numerical	Number of call failures
Complains	Categorical	0: No Complaint 1: Complaint
Charge Amount	Categorical	0: Lowest Amount 1: Highest Amount
Seconds of Use	Numerical	Total seconds of calls
Frequency of Use	Numerical	Total number of calls
Frequency of SMS	Numerical	Total number of text messages
Distant Call Numbers	Numerical	Total number of distinct calls
Tariff Plan	Categorical	1: Pay as you go 2: Contractual
Age Group	Categorical	1: Younger age 5: Older age
Status	Categorical	1: Active 2: Not active
Customer Value	Numerical	The calculated value of a customer
Subscription Length	Numerical	Total months of subscription
Churn	Categorical	0: Non-churn 1: Churn

B. EXPLORING THE DATA

1. Data Preprocessing

There may be several impurities in the dataset which needs to be removed before feeding it to the model. Data pre-processing aims to perform this and get rid of all these impurities.

Steps involved in Data Preprocessing:

a.) **Removing Garbage Values**

There can be a lot of noisy and useless elements in the raw data. We need to drop or replace them before training the model.

Eg: The attribute **'*complains*'** is a Boolean type, but also contains the following values: **-1, 11, -112, -11, 1999, 111111111**. All these values should be converted to Boolean i.e. 0 or 1.

b.) **Removing Null Values**

Another aspect of Data Cleaning is to treat the missing values in the dataset.

Some of the attributes, such as **'seconds_of_use'**, **'customer_value'** have a lot of null values which need to be either replaced or dropped.

Since some of the columns have outliers, we used **Median** to replace the null values instead of Mean - which are prone to outliers.

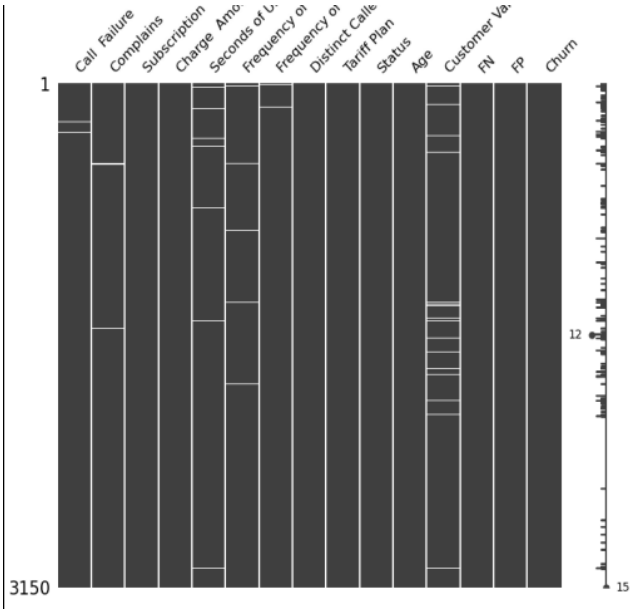

Fig. 1.1 Missing Values (Missingno Matrix Plot)

Observation: The white lines in the figure indicate the missing values in each column. Columns such as 'call_failure', 'complains', 'charge_amount', 'seconds_of_use', 'frequesncy_of_use' and 'customer_value' contain missing values, which needs to be treated.

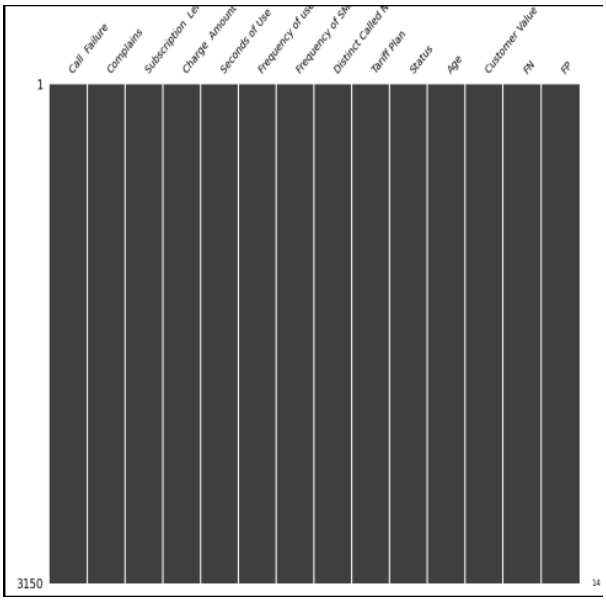

Fig. 1.2 No Missing Values (After preprocessing)

Observation: We can see from the graph above that there are no null values in the dataset after processing it.

c.) **Removing Outliers**

An outlier is a value that substantially differs from other values in a dataset and diverges from an overall pattern on a sample. They should be removed as most of them represent measurement errors, data entry or processing errors, or poor sampling.

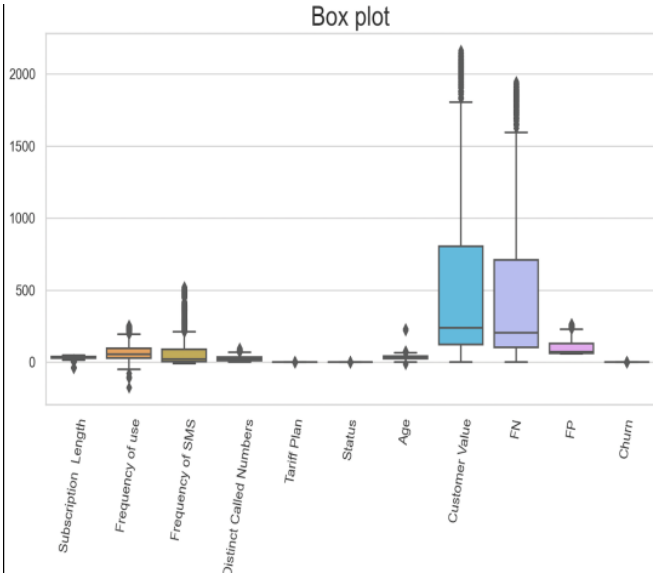

Fig. 2.1 Outlier Detection

Observation: As can be seen from *Fig. 2*, seconds_of_use has the most number of outliers.

Method used to find the outliers is the ***z-score*** method. It describes the position of a raw score in terms of its distance from the mean, when measured in standard deviation units.

Z-Score is essentially how many standard deviations away is the actual value from the mean value, you can define the

threshold value for the z score to classify a point as an outlier or not in the current scheme of things. Here we pass **4** for the threshold value.

Once the outliers were detected using the z-score method. They were then replaced with the median of the column.

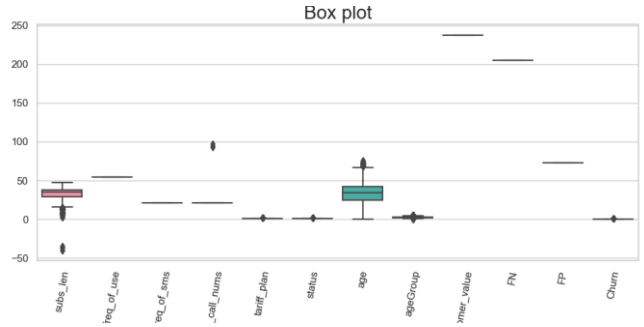

Fig. 2.2 Outlier Detection

Observation: We can see that the outliers have significantly decreased after pre-processing.

2. Data Visualization

It is significantly important to have a technique that will give us a better understanding of the dataset. This is where Data Visualization comes in handy. By using tools like **graphs**, **charts** and **maps**, data visualization provides an accessible way to identify outliers and different patterns in the dataset.

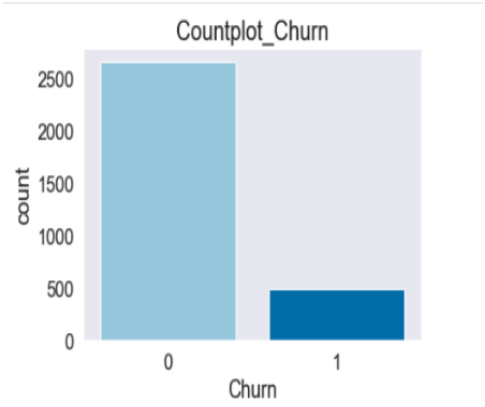

Fig. 3. Churn Distribution

***Observation*:** For the predictor feature, we observe that there are 84.29 percent non-churn customers and 15.71 percent churn customers in total, indicating a data imbalance.

Distribution of some categorical features:

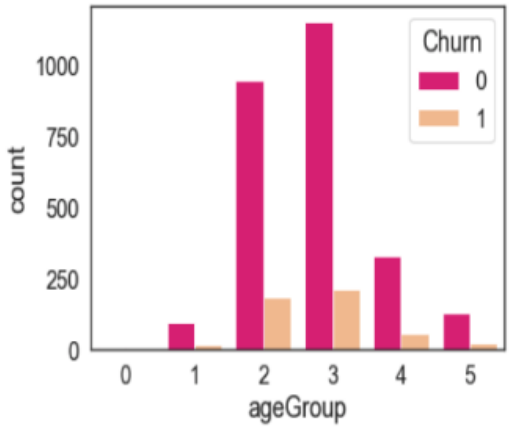

Fig. 4. Age Group Distribution

Observation: We notice that most number of the customers which are likely to churn are between the ages of 30-40 followed by age group 20-30 and over 40 years.

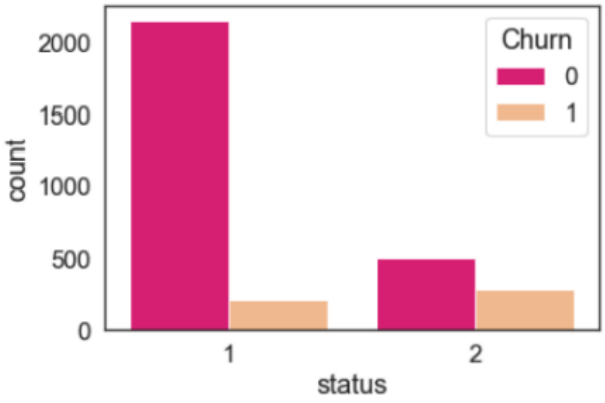

Figure 5: Frequency Distribution of Status feature.

Observation: We observe that the inactive customers are more likely to churn when compared to the customers which are actively using the service or subscription

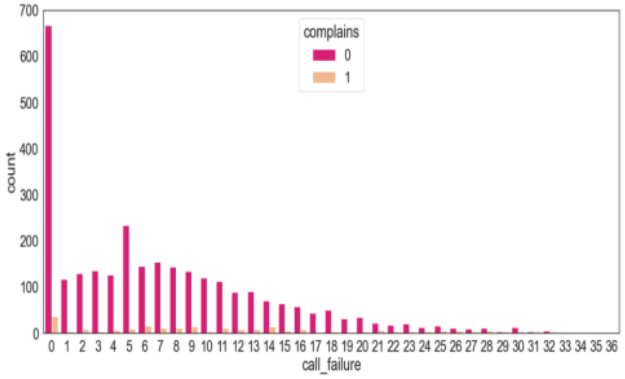

Figure 6 : Distribution of Call failure

Observation: From the fig. 6 we observe that there's a higher percentage of consumers who have no call failures and thus no complaints.

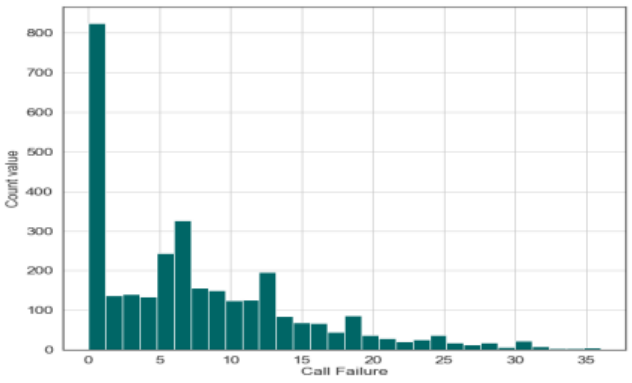

Figure 7 : Call Failure

Observation: We can see that approximately 2000 customers had call failures which shows that this factor might be decisive in calculating the customer churn rate.

3. Feature Selection

We used **SelectKBest** feature selection technique to select the top features to train different multi-classification model. We can visualize with the help of horizontal bar plot shown below.

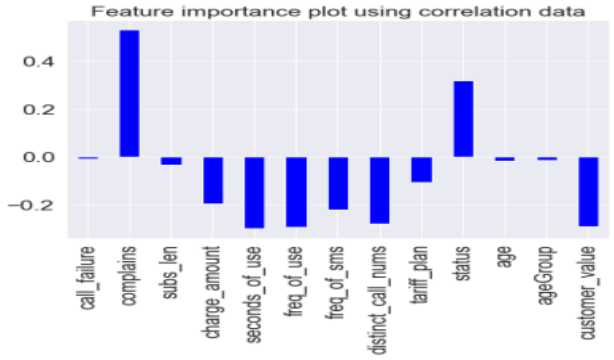

Fig. 8. Feature Selection Graph

Observation*: From the graph we observe that status and complain are influential features for this dataset.*

II. METHODS

We chose three models to train and test our preprocessed dataset: Support Vector Classifier (SVC), Decision Tree, and Logistic Regression.

1. Support Vector Classifier (SVC)

It is a linear model that can be used to solve classification and regression problems. It can solve both linear and nonlinear problems. The algorithm generates a line or hyper-plane that divides the data into categories. Support Vector Classifier yields following results:

Accuracy	AUC	Precision	Recall	F1-Score
84.3	50	71	84.3	77.1

Fig. 9. SVC Results

2. Decision Tree Classifier

It is supervised machine learning that categorizes or predicts outcomes based on the answers to a previous set of questions. Decision Tree yields following results:

Accuracy	AUC	Precision	Recall	F1-Score
91.1	82.4	90.9	91.1	91

Fig. 10. Decision Tree Classifier

3. Logistic Regression

Logistic regression is an example of supervised learning. It is used to calculate or predict the probability of a binary (yes/no) event occurring. An example of logistic regression could be applying machine learning to determine if a person is likely to be infected with COVID-19 or not.

Accuracy	AUC	Precision	Recall	F1-Score
91.1	75	90.8	91.1	90.1

Fig. 11. Logistic Regression Classifier Results

III. PERFORMANCE METRICS

1) **Precision**: The number of positive class predictions that actually belong to the positive class is measured by precision.

2) **Recall**: The number of positive class predictions that actually belong to the positive class is measured by precision.

3) **F1-Score**: F1-Score generates a single score that accounts for both precision and recall concerns in a single number.

4) **Accuracy**: It's the proportion of correct predictions to total input samples. It only works if each class has an equal amount of samples.

5) **AUC**: The Area Under the Curve (AUC) is a curve that measures a classifier's ability to distinguish between classes. The greater the AUC, the better.

Note: Since our data is class imbalanced, we majorly rely on F1-Score, Recall and Precision.

IV. CONCLUSION

In this analysis we experienced three prominent classification techniques using an Iranian telecommunication company dataset. Decision Tree significantly outperformed the other classifiers. This study also investigated that complaints and customer status are the major factors that influence customer churn significantly. Based on these factors, we suggest potential approaches that will enable the telecom company reduce the customer attrition rate considerably.

Approaches for using customer status as an alarm to churn potential:

a) Monitor the change in customer status as an alarm to churn potential.

b) Provide special offers and services to customer with inactive status.

c) Identify factors that make the customer status inactive and try to avoid them.

Approaches to avoiding customer churn due to dissatisfaction:

a) Conduct direct and indirect polling to determine customer expectations and perceptions about operator services.

b) Consider programs for rewarding long-term customers as lucrative assets of the organization.

c) Try to improve network coverage.

LIMITATIONS AND FUTURE RESEARCH

Lack of access to different types of data is the main limitation of this study, for example service costs, geographic location, types of service(phone/internet), dependence of customer recount as an important factor of customer churn. Having multiple service providers data could be very useful for understanding customer retention behaviour more thoroughly. Overcoming these limitations can be done in future research. Also, customer probable churn time could be considered. Using time series methods can be useful in extracting churn prediction function and calculating customer churn probability in certain time interval

REFERENCES

[1] https://analyticsindiamag.com/tips-for-automating-edausing-pandas-profiling-sweetviz-and-autoviz-in-python/J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] https://tinyurl.com/TelecomCustomerChurnDataset K. Elissa, “Title of paper if known,” unpublished.

[3] Ahmed U, Khan A, Khan SH, Basit A, Haq IU, Lee YS (2019) Transfer learning and meta classification based deep churn prediction system for telecom industry..

[4] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A, Hussain A (2016) Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access 4:7940–7957

[5] https://techsee.me/blog/telecom-customer-retention