**Project report on Customer Churn Analysis**

**Submitted towards partial fulfilment of the criteria**

**for award of PGPDSE by Great Lakes Institute of Management**

**Submitted By**

**Group No. 9 [Batch: September - 2019]**

**Group Members**

1. **Ashwin. V**
2. **Prasad. A**
3. **Shintu Mathew**
4. **Sreelalan. S**
5. **Vigneshwaran**

**Research Supervisor**

**Mr. Srikar Muppidi**



**Great Lakes Institute of Management**

**ABSTRACT**

Churn tells us the number of customers who have unsubscribed or canceled their service contract with an organization. Customers who reject the service offered is always a serious issue that must be looked into. In order to reduce the churn, the company have to take necessary steps. The aim of this research is to build a model that identifies the customers who might switch their telecom service provider soon. By designing this model, we hope to classify the customers based on their probability of churning and provide the company with necessary insights so that the company can do the necessary steps to retain the customers. Technique used here is predictive modelling.

* Tools: (Python, Tableau, Jupyter notebook)
* Domain: Telecom

**Problem Statement:**

The company is having a churn rate of 26%. The company wants to reduce the churn rate and increase the profitability. The company wants to identify the potential churners so that the precautionary measures can be taken up front so that the valuable customers can be retained. The company also want to identify the potential causes for the previous churns so that the company can work on specific areas for the improvement.

**PROPOSED SOLUTION:**

The steps we are going to implement to deal with the problem,

* Customer Churn Analysis

This step is understanding the scenario prevailing and how to change that to desired outcome. Here the objective is reducing customer churn by identifying potential churn candidates beforehand, and take proactive actions to make them stay.

* Data loading & Cleaning

Following is the data loading and cleansing the data sets. Here we deal with the missing values, dropping the columns, changing the index of the data frame, renaming the column and skipping the rows etc. also for feature selection or engineering.

* Exploratory data analysis & Data visualization

Exploratory Data Analysis is an initial process of analysis, in which you can summarize characteristics of data such as pattern, trends, outliers, and hypothesis testing using descriptive statistics and visualization. We start with checking out how our data looks like and visualize how it interacts with our label churned or not.

* Statistical approach

Statistics is a collection of tools that gives answers to important questions about data. We can use statistical methods to transform raw observations into information. We can use inferential statistical methods to reason from small samples of data to whole domains.

* Data Preprocessing

After the data going through all the above steps, we can train our machine learning model. But first, we need to separate the variable that we’re predicting from the dataset as train and test. So we can later evaluate the performance of our machine learning model.

* Building prediction model

To make good predictions, we firstly need to find the right model and secondly need to evaluate that the algorithm actually works. This step takes a few iterations, we will keep this simple and stop as soon as the results fit our needs.

* Evaluating model performance

Here, we will get a classification rate, considered to get good or bad accuracy. Precision score is about being precise, i.e., how precise your model is. In other words, you can say, when a model makes a prediction, how often it is correct.

* Conclusion

Here we choose the best model that gives better accuracy. Followed by this required actions will be taken.

**EXPLORATORY DATA ANALYSIS:**

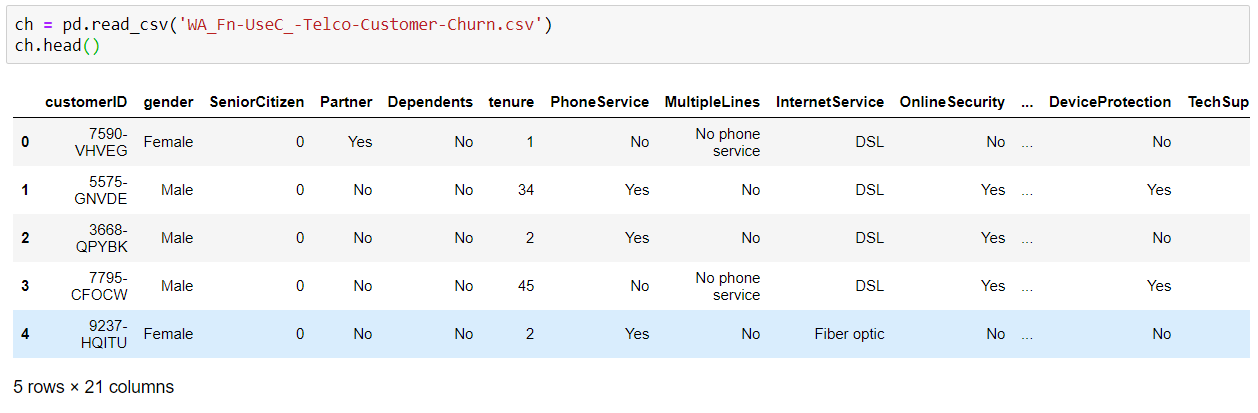
****

Fig 1: Telecom Dataset

Telecom dataset with 7043 rows and 21 columns.

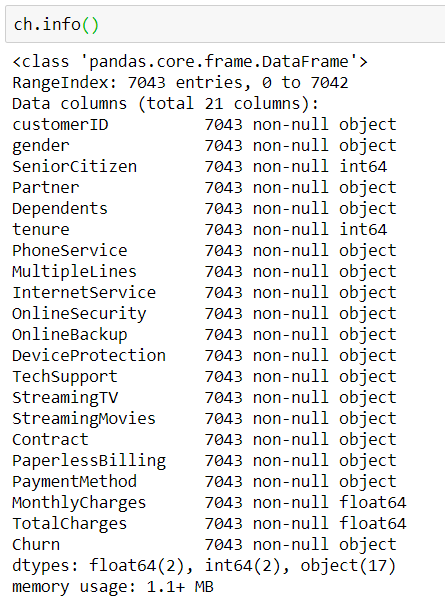


Fig 2: Dataset Summary

The summary includes list of all columns with their data types and the number of non-null values in each column. We also have the value of range index provided for the index axis.

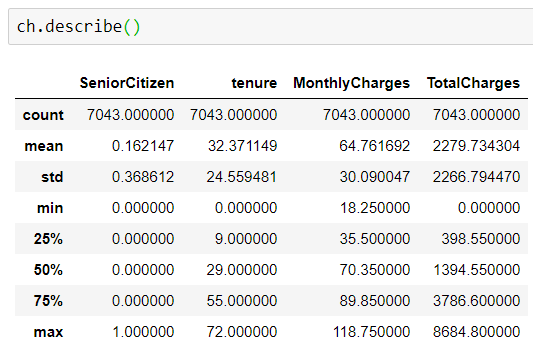


Fig 3: Statistical Information

As shown in the output image, Statistical description of data frame was returned with the respective passed percentiles. Since there are no missing values, we did not get nay NaN.

**Explanation of each feature:**

* Customer ID: Denotes the Customer ID.
* Gender: Whether the customer is a male or a female.
* Senior Citizen: Whether the customer is a senior citizen or not (1, 0)
* Partner: Whether the customer has a partner or not (Yes, No)
* Dependents: Whether the customer has dependents or not (Yes, No)
* Tenure: Number of months the customer has stayed with the company.
* 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, No phone service)
* Internet Service: Customer’s internet service provider (DSL, Fiber optic, No)
* Online Security: Whether the customer has online security or not (Yes, No, No internet service)
* Online Backup: Whether the customer has online backup or not (Yes, No, No internet service)
* Device Protection: Whether the customer has device protection or not (Yes, No, No internet service)
* Tech Support: Whether the customer has tech support or not (Yes, No, No internet service)
* Streaming TV: Whether the customer has streaming TV or not (Yes, No, No internet service)
* Streaming Movies: Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract: The contract term of the customer (Month-to-month, One year, Two year)
* Paperless Billing: Whether the customer has paperless billing or not (Yes, No)
* Payment Method: The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
* Monthly Charges: The amount charged to the customer monthly
* Total Charges: The total amount charged to the customer
* Churn: Whether the customer churned or not (Yes or No)

**PRILIMINARY EDA:**

The preliminary eda consists of the number of counts of each variables in a feature.

|  |  |
| --- | --- |
| **Churn**  No 5174  Yes 1869 | **Dependents**    No 4933  Yes 2110 |
| **Gender**  Male 3555  Female 3488 | **Partner**  No 3641  Yes 3402 |
| **Phone Service**    Yes 6361  No 682 | **Senior Citizen**    0 5901  1 1142 |
| **Gender Vs Churn**  Churn No Yes  gender  Female 2549 939  Male 2625 930 | **Phone Service Vs Churn**  Churn No Yes  PhoneService  No 512 170  Yes 4662 1699 |
| **Churn Vs Contract**    Churn No Yes  Contract  Month-to-month 2220 1655  One year 1307 166  Two year 1647 48 | **Churn Vs Internet Service**    Churn No Yes  InternetService  DSL 1962 459  Fiber optic 1799 1297  No 1413 113 |
| **Churn Vs paperless Billing**    Churn No Yes  PaperlessBilling  No 2403 469  Yes 2771 1400 | **Churn vs PaymentMethod**  Churn No Yes  PaymentMethod  Bank transfer 1286 258  Credit card 1290 232  Electronic check 1294 1071  Mailed check 1304 308 |

**EXPLORATORY VISUALIZATION**

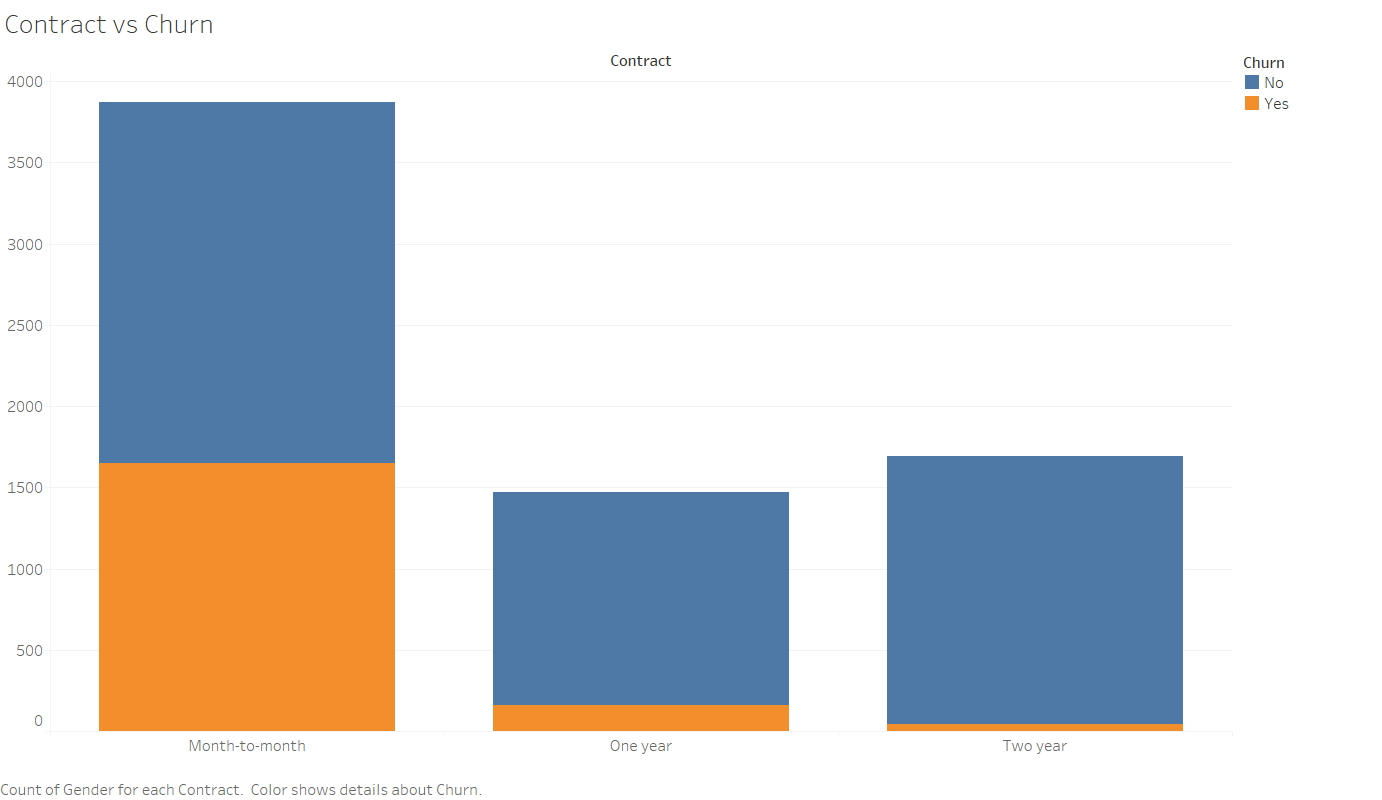


Fig 4: Contract vs Churn

Of all contracts, month to month contract churn ratio as well as not being churned is more. Customers who opted one year and two year contract are likely to be satisfied and staying with the company.

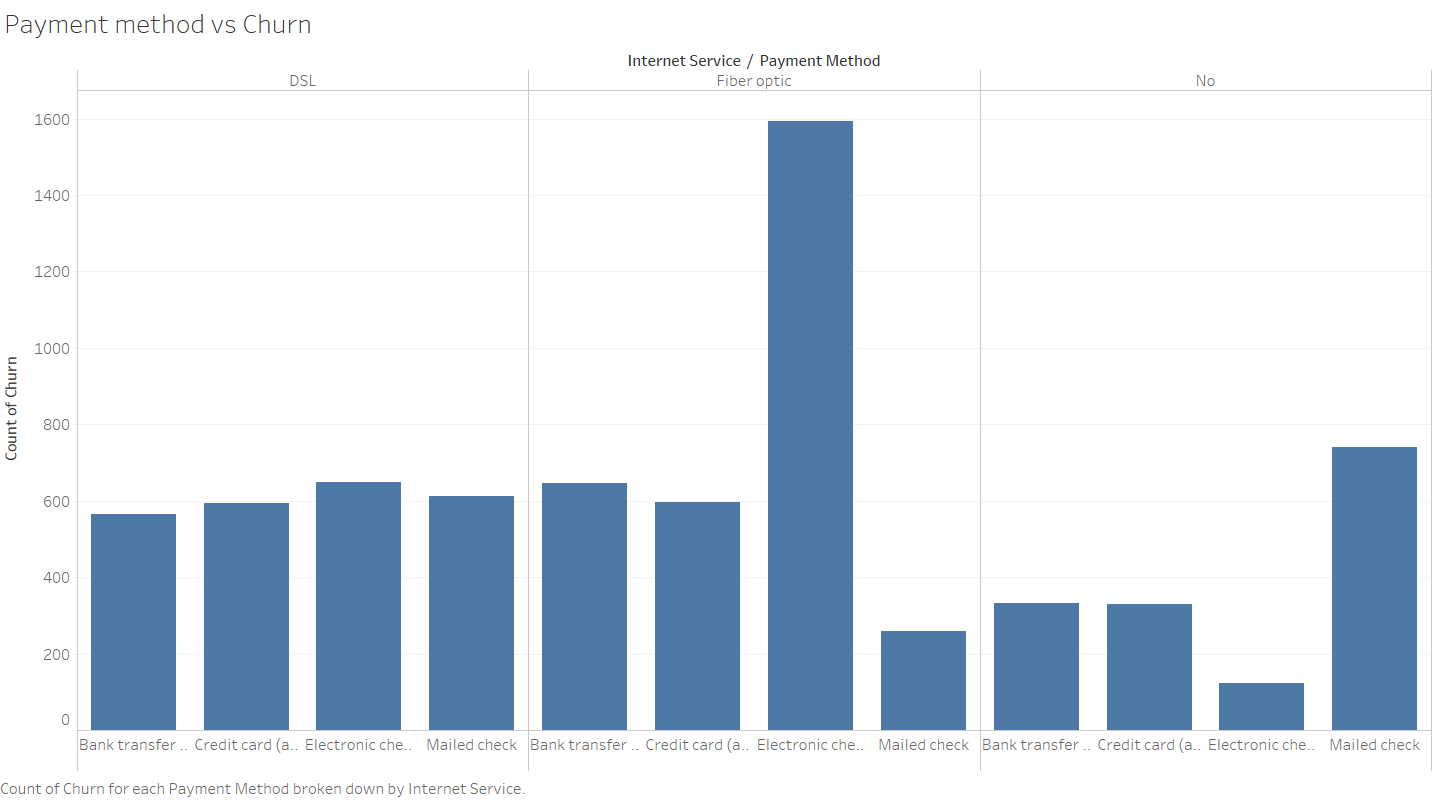


Fig 5: Payment method vs Churn

From this, we can see that customers opted the services are paying through all 4 kinds payment methods, yet the payment through electronic check is high as it can be processed in fewer steps, for fibre optic service. Customers with no services and their payment count is too low among others but for mailed checks.

|  |  |
| --- | --- |
|  |  |

Fig 6: Internet Service vs Customers

None senior citizen is high in numbers in all three services and the churn ratio is also considerably low except for the optic fibre. Even the churn ratio of senior citizens is very low since they are bound with technical features.

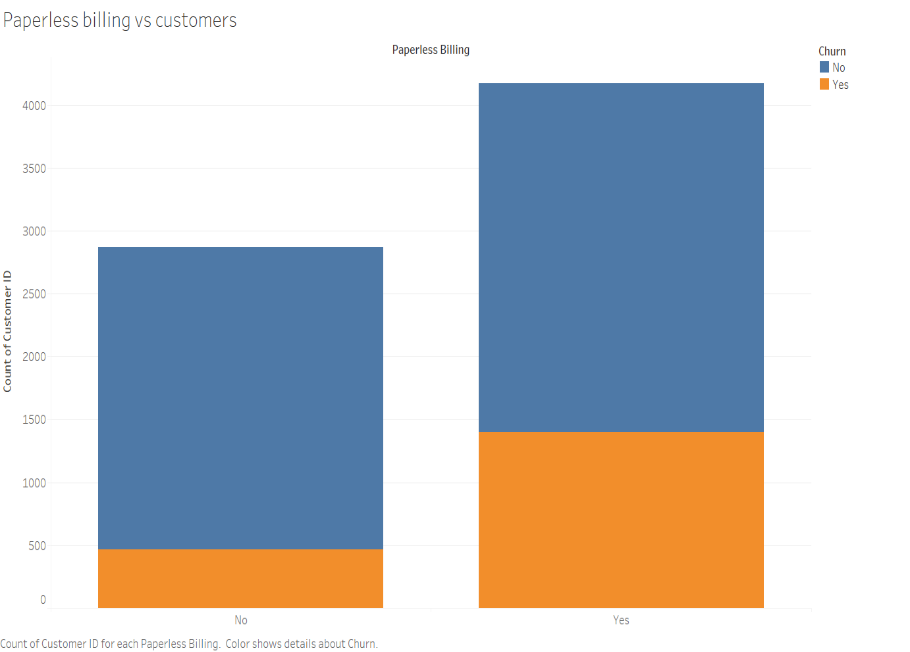


Fig 7: Paperless billing vs customers

Customers saying no to paperless billing and its churn ratio is very low than paperless transaction. It is the none senior citizens whose count is high in both paperless and traditional billing.

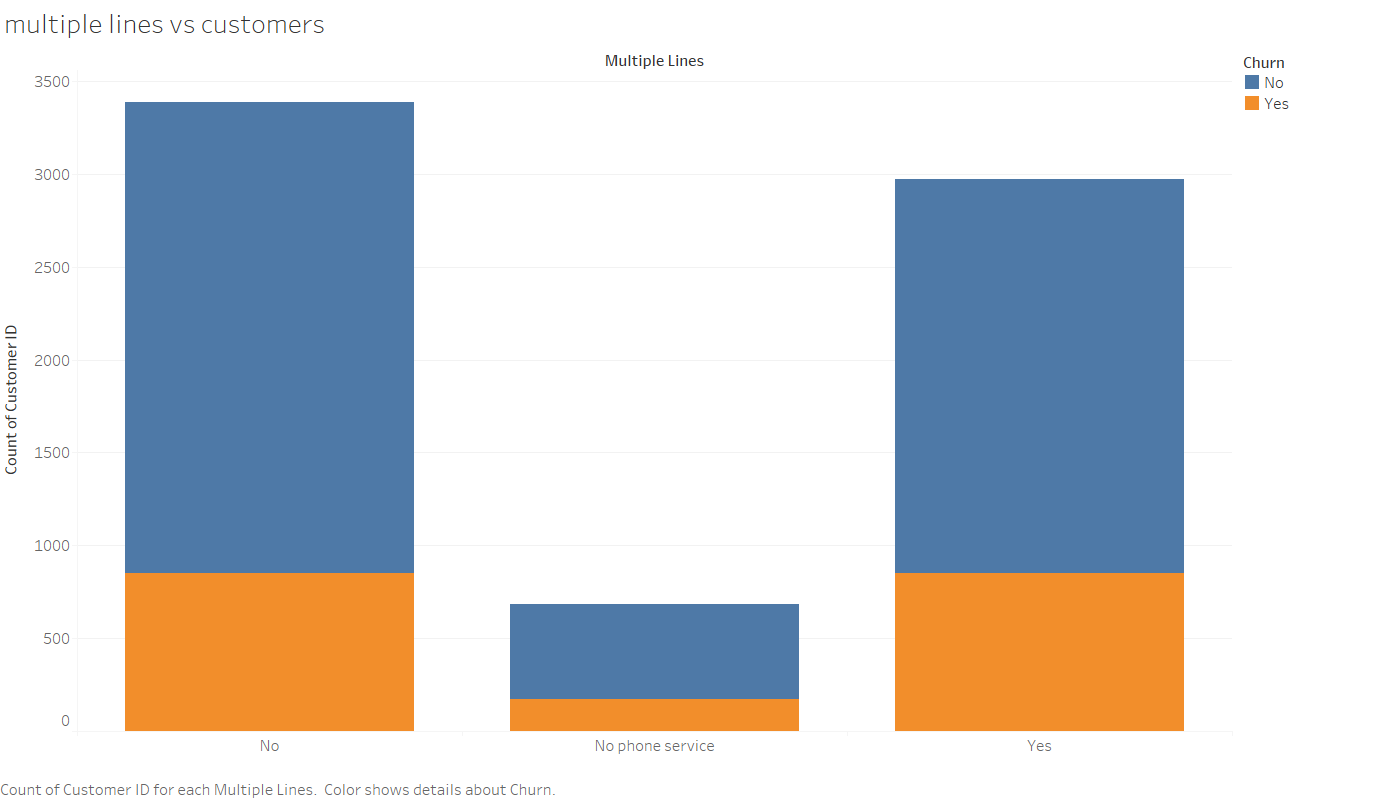


Fig 8: Multiple lines vs Customers

Multiple lines users count are decent in numbers but not more than users with no multiple lines. Even the churn ratio in both no & yes lies common. So the churn does its role the same even if the customers’ possess single or multiple lines.

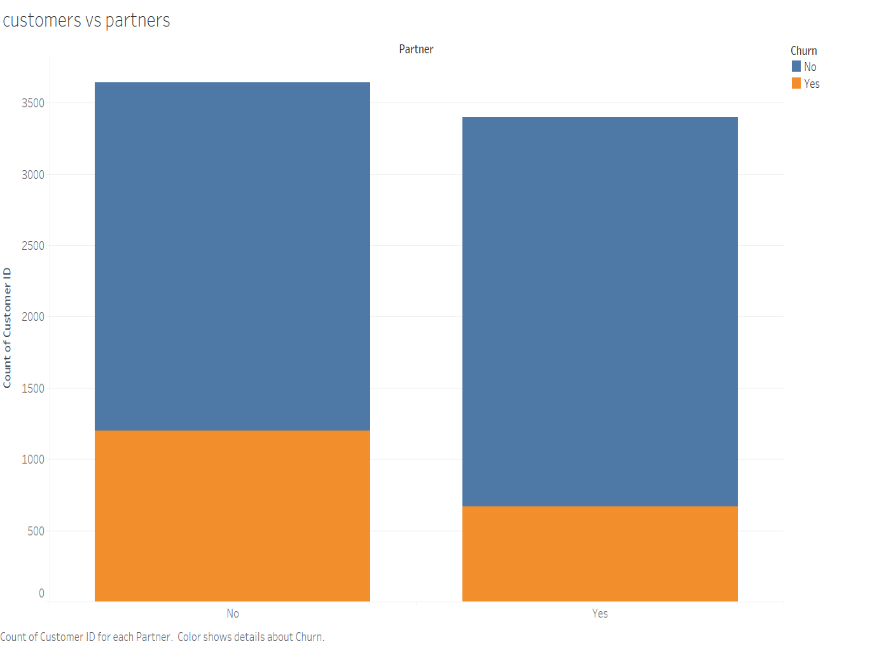
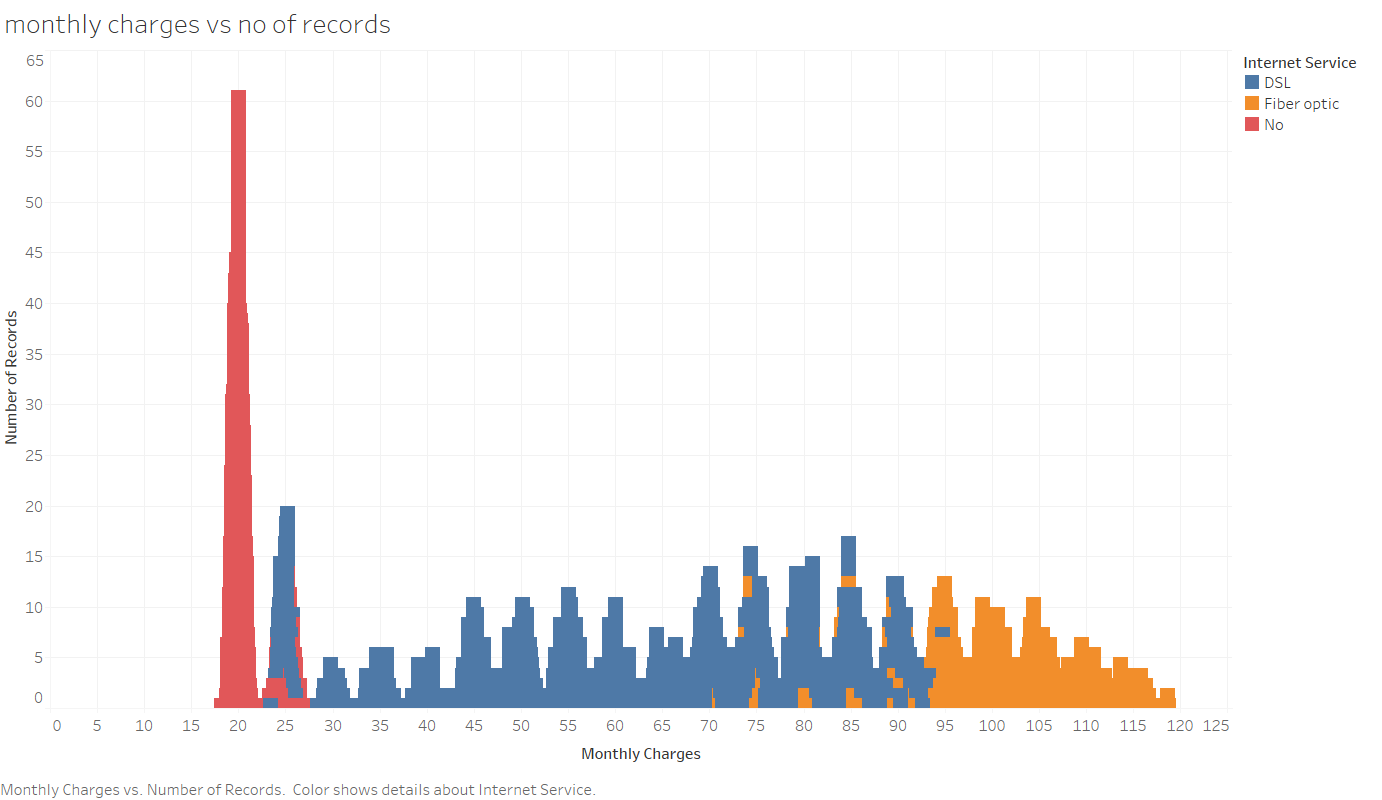


Fig 9: Customers vs Partners

Here customers with partners churn rate is less than the customers without partners whose count is more in numbers in opting the service also possessing higher churn rate.

Fig 10: Monthly Charges

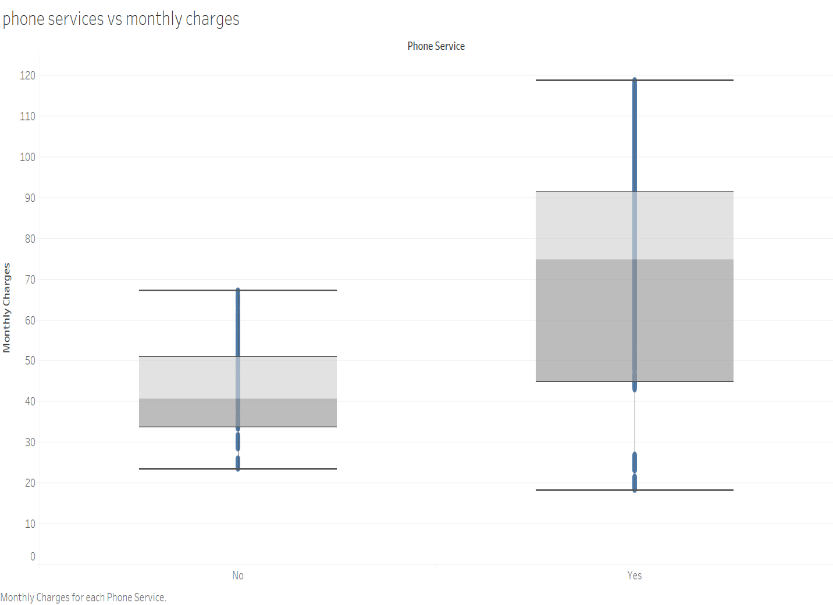


Fig 11: Box plot of phone services vs monthly charges

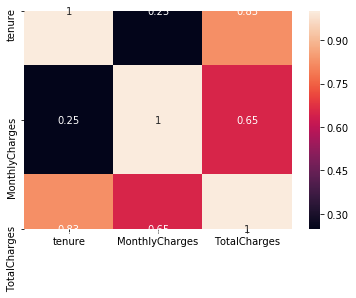


Fig 12: Heatmap of the continuous variables pointing to the correlation between variables

**Statistical Significance**

From the above EDA, we understood that certain features have dependency depending upon the churn. So, in order to check the variations, we can use statististical methods to check the significance of the target variable with respect to the rest of the features.

We propose two hypotheses to test the significance of the features with respect to churn.

H0 is the null hypothesis and Ha is alternate hypothesis.

**H0**- Churn does not depend upon the feature.

**Ha**- Churn depends upon the feature.

Significance threshold (alpha – 0.05)

To check if the features are significant in contributing to the churn, we use Chi square test to verify the proposed hypothesis as the majority of the features are categorical. To check the continuous variables, we will make use of z test. After using the chi square test, we will get a p value. If the P values is bigger than alpha which is 0.05, then we accept the null hypothesis, else we reject the null hypothesis. After performing chi square test, We got the results as below

**Gender vs Churn** : 0.48657873605618596

**Senior Citizen vs Churn**: 1.510066805092378e-36

**Partner vs Churn**: 2.1399113440759935e-36

**Dependents vs Churn**: 4.9249216612154196e-43

**Phone Service vs Churn**: 0.3387825358066928

**Multiple Lines vs Churn**: 0.0034643829548773

**Internet Service vs Churn**: 9.571788222840544e-160

**Online Security vs Churn**: 2.661149635176552e-185

**Online Backup vs Churn**: 2.0797592160864276e-131

**Device Protection vs Churn**: 5.505219496457244e-122

**Tech Support vs Churn**: 1.4430840279998987e-180

**Streaming TV vs Churn**: 5.528994485739183e-82

**Streaming Movies vs Churn**: 2.667756755723681e-82

**Contract vs Churn**: 5.863038300673391e-258

**Paperless Billing vs Churn**: 4.073354668665985e-58

**Payment Method vs Churn**: 3.6823546520097993e-140

From the above findings, we can understand that the p value of Gender is greater than the alpha 0.05. So, in case of gender we accept the null hypothesis. Churn is independent of gender. But for rest of the features, the churn is affected in one way or other. So, for rest of the features, we reject the null hypothesis. So, we can assume that the rest of the features contribute to the churn to an extent.

Now checking the statistical significance of continuous variables, we make use of z test. Here values are divided from same feature depending upon the churn.

H0 is the null hypothesis and Ha is alternate hypothesis.

**H0**- v0 = v1

**Ha**- v0 != v1

Significance threshold (alpha – 0.05)

Where v0 is the set of variables where churn is 0 and v1 is the set of variables where the churn is 1. To check if the features are significant in contributing to the churn, we use z test to verify the proposed hypothesis After using the z test, we will get a p value. If the P values is bigger than alpha which is 0.05, then we accept the null hypothesis, else we reject the null hypothesis. After performing z test, We got the results as below

**Tenure** :7.048598128156636e-219

**Monthly Charges** : 1.9953885912944285e-61

**Total Charges** :1.1791224704797705e-64

From the above findings, it is understood that all the P values are smaller than the alpha 0.05. So, for all 3 variables, we reject null hypothesis. So, for all three variables, -v0 != v1. With this finding, we proceed with building a rough model.

**A Base Model Building**

We are building a sample model using Random forest classifier. The reasons for picking random classifier are

* Random forest classifier can handle large amount of data and it will try to maintain higher accuracy always.
* Random forest classifier will handle the missing values potently.
* Random forest will avoid overfitting.
* Random forest handles larger dimensions efficiently.

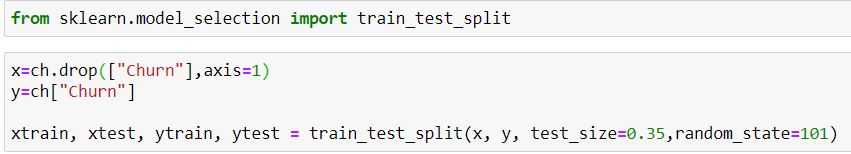


Fig 13: Splitting of train, test data

All kinds of categorical data are converted using LabelEncoding. After converting categorical data, all continuous data are transformed using any kind of transformation. In the problem we have used standard scalar. After scaling the data, using the train test split, we are splitting the data into train data and test data. As this is a sample model, we are retaining all the columns except customer id. Churn is taken as Y and rest of the data are taken as X. The test size is set as 0.35 and random state is set as 101. The random state is set so that the same split of data can be used in future models also. After fitting the model, the y values are predicted using predict function. Using confusion matrix, we can understand the efficiency of the base model. Using classification report and confusion matrix, following data is obtained.

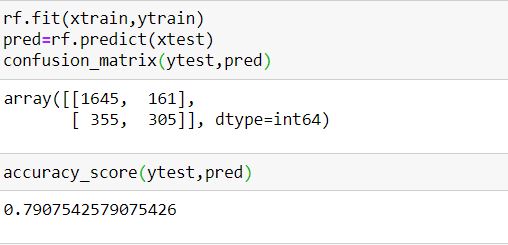


Fig 14: Confusion matrix of the base model

With the base model, in the confusion matrix, we obtained

* True positive = 305
* False Positive = 355
* True negative = 1645
* False negative = 161

Using the classification, when we calculate the different metrics of classification matrix, we obtain

* Accuracy = 0.7907
* Miss classification = 0.2092
* Recall in terms of (1) = 0.46
* Recall in terms of (0)=0.91
* Precision in terms of (1)=0.65
* Precision in terms of (0)=0.82

From the base model, it is evident that, the model has a good accuracy of 0.7907. But When we closely examine, we can understand that true negative are correctly classified with a recall value of 0.91 but the true positive values are classified with a recall value of only 0.46. The precision of 0 is 0.82 while the precision of 1 is only 0.65. The problem statement clearly mentions that the company needs to identify the people who is potential to churn. But the model in the current state can only predict the people who will not churn in the future. Also, the model has misclassified the churners at a higher rate. This will masquerade churners into non churners rendering impossible for us to identify the churners exactly. So, the future models has to be modelled in such a way that the recall and precision of churners is increased so that the model can identy the churners and potential churners effectively.

**Future roadmap of the project**

* Build model to identify churners
* The model should identify the potential churners
* Various algorithms should be implemented as a part of improvement of the model
* Various parametric tuning, feature engineering methods are implemented to improve the accuracy of the model
* Tuning the model in such a way that, the model should work efficiently when new type of data previously unknown to the model is introduced.