Telecom Customer Churn Prediction

Introduction:

Customer churn, also known as customer attrition, customer turnover, or customer defection, is the loss of clients or customers. Telephone service companies, e-commerce companies, internet service providers, pay TV companies, insurance firms, etc., often use customer churn analysis and customer churn rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Churn rate is the amount of customers or subscribers who cut ties with the service or company during a given time period. These customers have "churned".

Problem at hand:

Predicting the behaviour of the customers leading to churn.

Value to client:

A telecom company has been affected by the increasing number of customers subscribing to the services of a competitor. It is much more expensive to attract new customers than retaining old customers. At the same time, spending too much on or spending on the wrong factor for retaining a customer who has no intention to leave (or who was not leaving for that factor which was addressed) could be a waste of money. Therefore it is important to identify the customer who has high probability of leaving. An analysis of the past records of the customers can give great insights on who might leave and what is the cause.

We can predict behaviour to retain customers. We can analyze all relevant customer data and develop focused customer retention programs.

Data Source:

www.kaggle.com/blastchar/telco-customer-churn

Methodology:

- Data Wrangling
- Exploratory Data Analysis (EDA) and Visualization
- Data Storytelling
- Training and Testing Machine Learning models
- Recommendations to retain the customers
- Scope for future work

Data Wrangling

The data was downloaded from Kaggle. Each row represents a customer and each column contains customer's attributes described on the column Metadata. The "**Churn**" column is our target.

The data set includes information about:

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

The data was then imported into a Pandas DataFrame for ease of data manipulation. The DataFrame contains 7043 rows (customers) and 21 columns (features).

Initially on checking the information of the DataFrame, no nulls were shown. On further inspection I saw the column 'TotalCharges' had a few rows with spaces, 11 rows to be exact. Tenure for all these 11 exact rows is also 0, churn is "No" as well. One can interpret this as belonging to all new customers, if indeed these are the only rows with tenure = 0 too. Hence we can set 'TotalCharges' to 0, whenever tenure is 0. So I used the replace function to convert the spaces to 0. I also checked for the unique values and items of each column. The 'customerID' column is of no use to us, so I created a new DataFrame without the 'customerID' column using the iloc method. The new shape of the DataFrame is 7043 rows by 20 columns.

On the whole this data was very clean and required minimal data cleaning steps. As of now the DataFrame looks good to be used for further analysis. If any data manipulation is required in the later part of the project, it'll be done accordingly.

Exploratory Data Analysis

Data Story

Here we are going to ask and answer a few questions about the data with respect to our dependent variable and also visualize & get insights from it.

Target (Dependent) Variable: Churn

Feature (Independent) Variables: 19 variables out of 20.

We can divide the independent variables into:

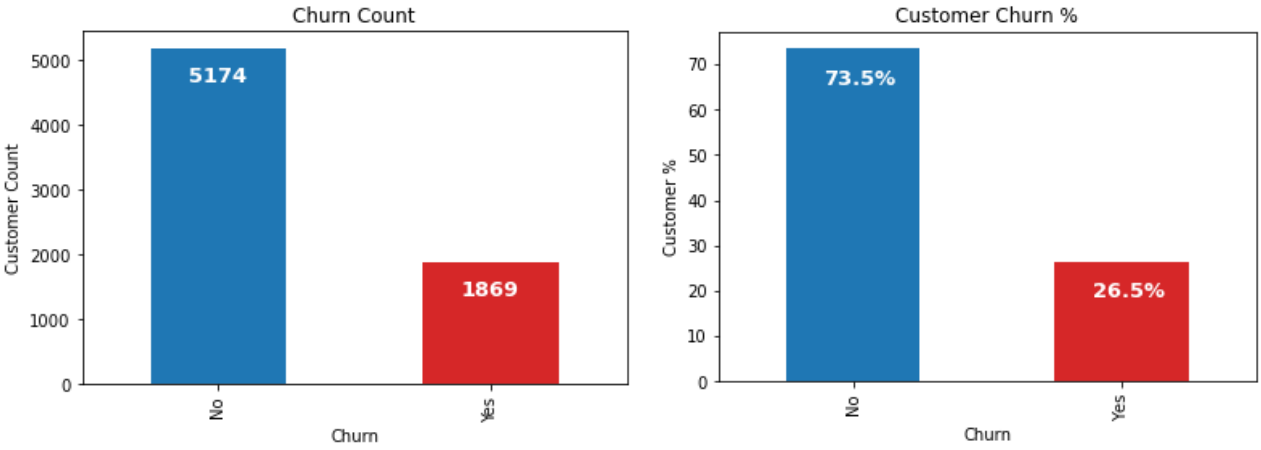
- **Person specific :** gender, SeniorCitizen, Partner, Dependents, tenure
- **Service specific :**
    - Phone : PhoneService, MultipleLines
    - Internet : InternetService, OnlineSecurity, OnlineBackup, StreamingTV, Streaming Movies, TechSupport, DeviceProtection
- **Money specific :** MonthlyCharges, TotalCharges, Contract, PaperlessBilling, PaymentMethod

The questions to which we seek answers:

- Is there a gender specific to churn?
- Are there any person specific trends in churn?
- Is there a correlation between tenure and churn?
- Is there a correlation between certain types of services and churn?
- Is there a correlation between different types of contract and churn?
- Is there a correlation between paperless billing and churn?
- Is there a correlation between different types of payment methods and churn?
- Is there a correlation between monthly charges and churn?
- Is there a correlation between total charges and churn?
- Is there a correlation between monthly charges and total charge with respect to churn?
- Is there a correlation between monthly charges and tenure with respect to churn?
- Is there a correlation between total charges and tenure with respect to churn?

Customer Churn in the data :

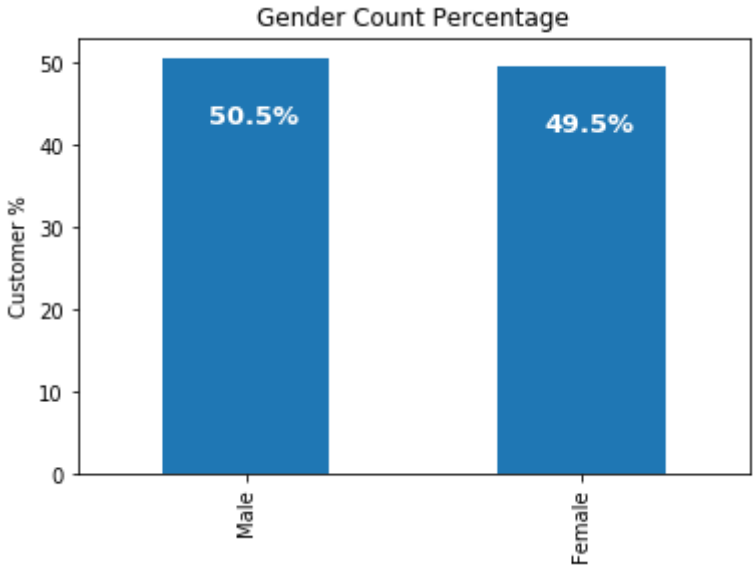
First let us check out the number of customers who have churned or not churned.



- **So, we can see from above:**
  - **Churn No - 5174 or 73.5%**
  - **Churn Yes - 1869 or 26.5%**

Gender distribution :

Let's see the gender distribution in the dataset.

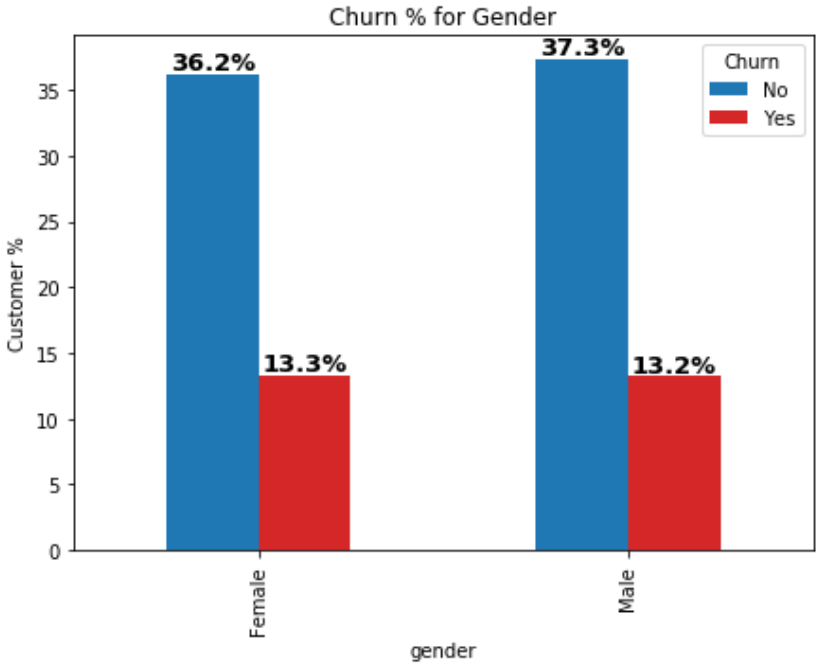


➢ We can see that the gender distribution looks balanced.

Visualizing the customer's attributes with respect to Churn :

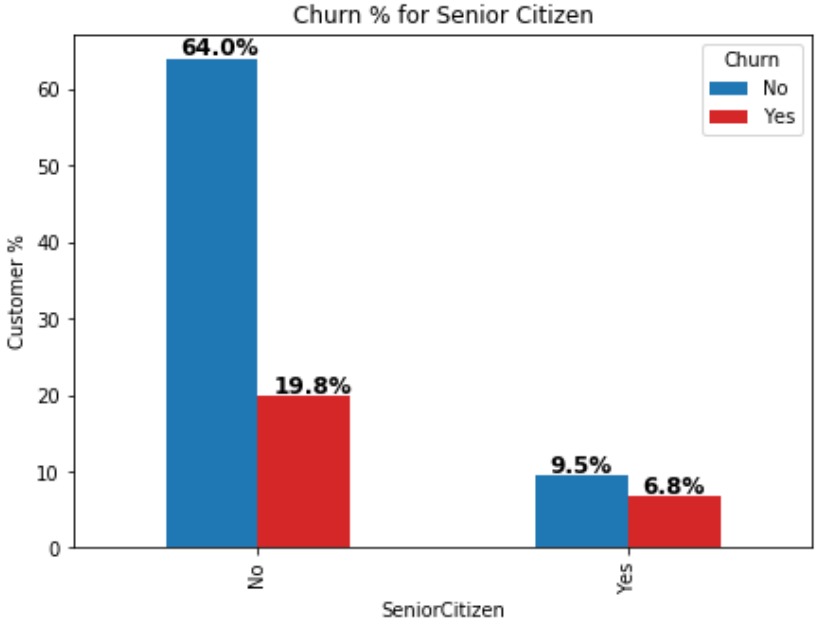
1. **Person specific attributes -**

Gender



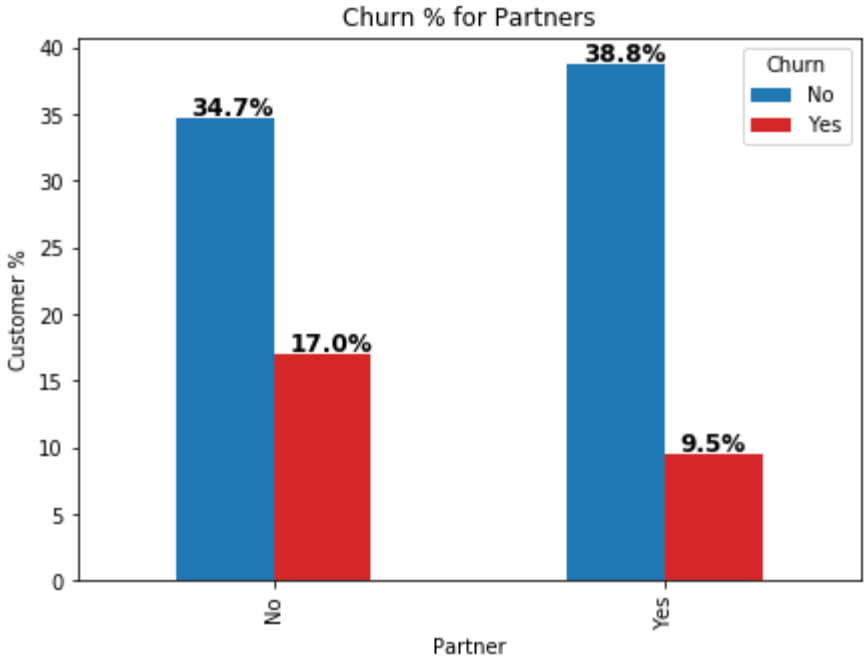
- The churn rate is not affected by the gender.

Senior Citizen



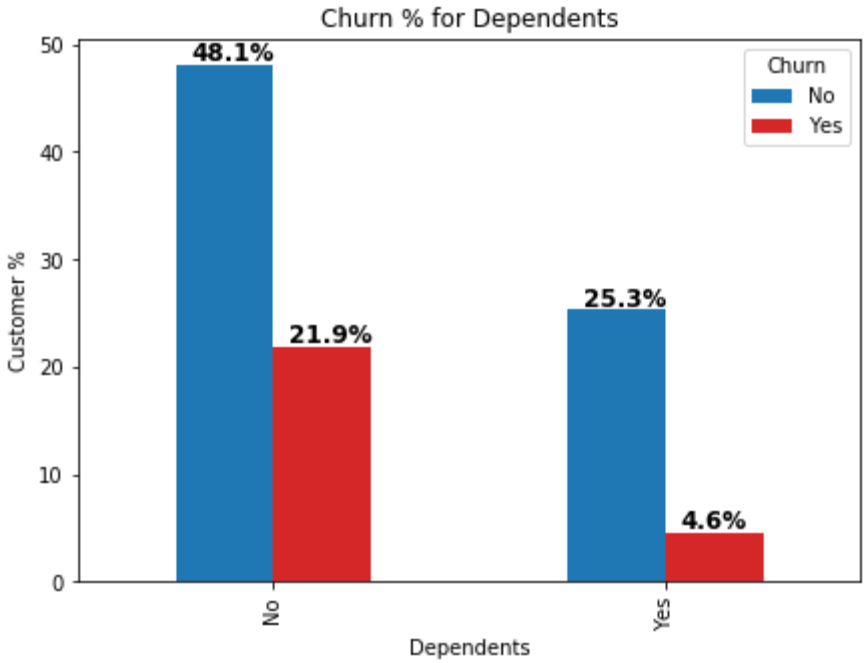
- Senior Citizens tend to churn less compared to non-senior citizens.

Partners



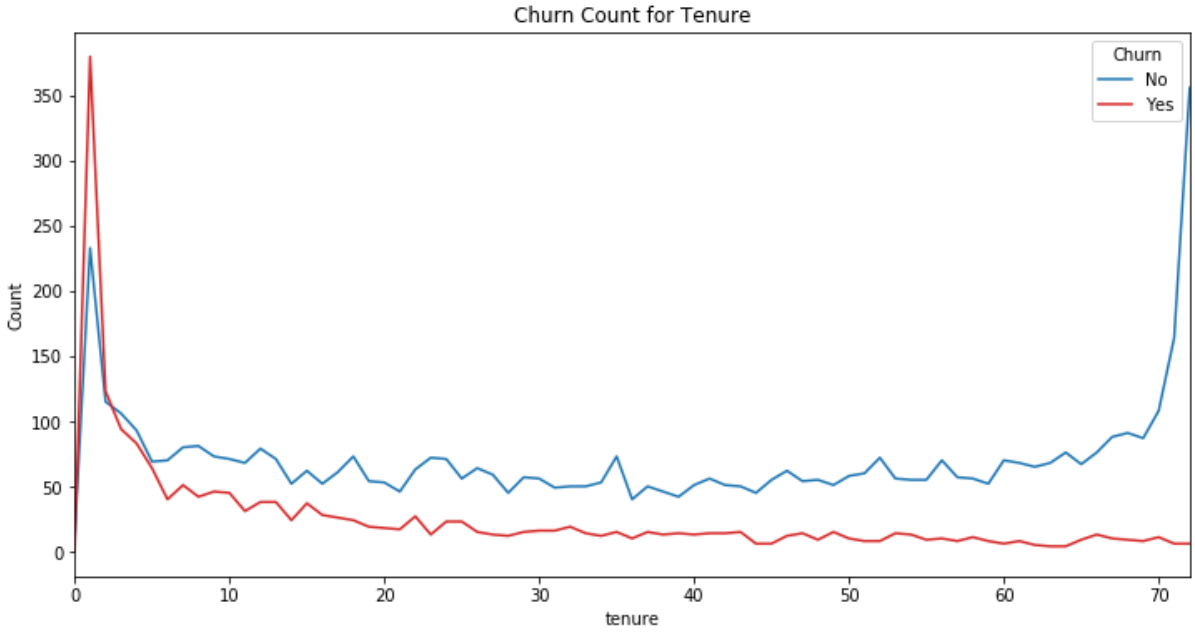
- Customers who don't have partners have higher churn rate.

Dependents



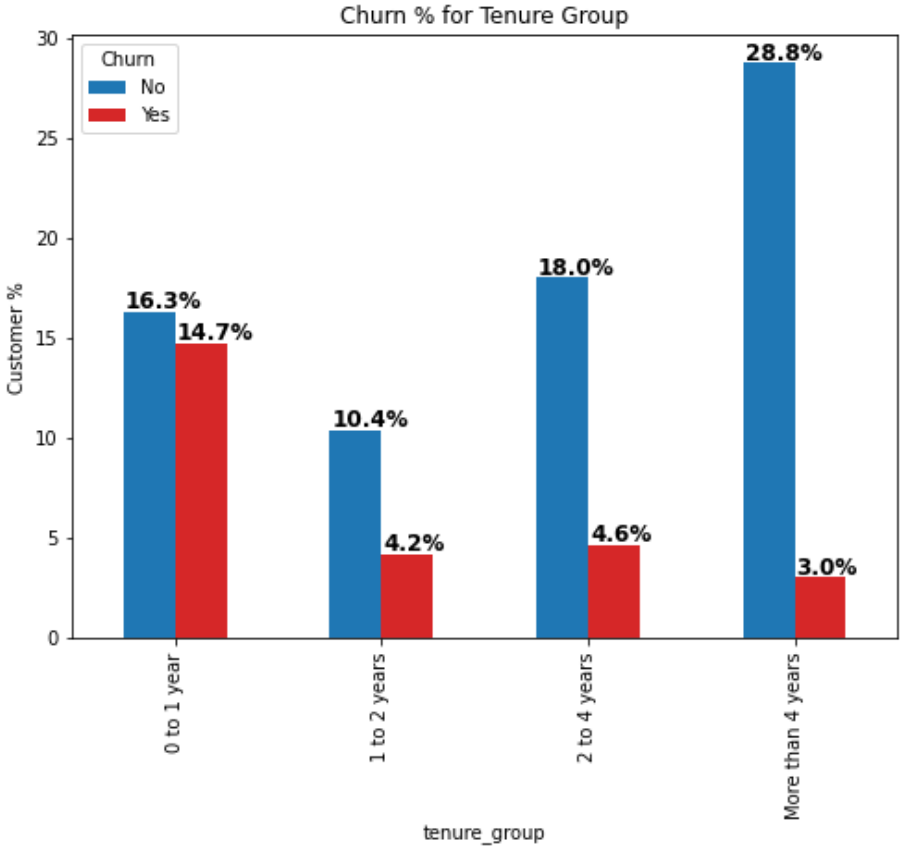
- Customers without dependents have higher churn rate.

Tenure



- Churn count decreases as the tenure increases.
- Customers tend to churn within the first few months or within a year.

Let's create 4 tenure groups to check the churn rate more clearly.

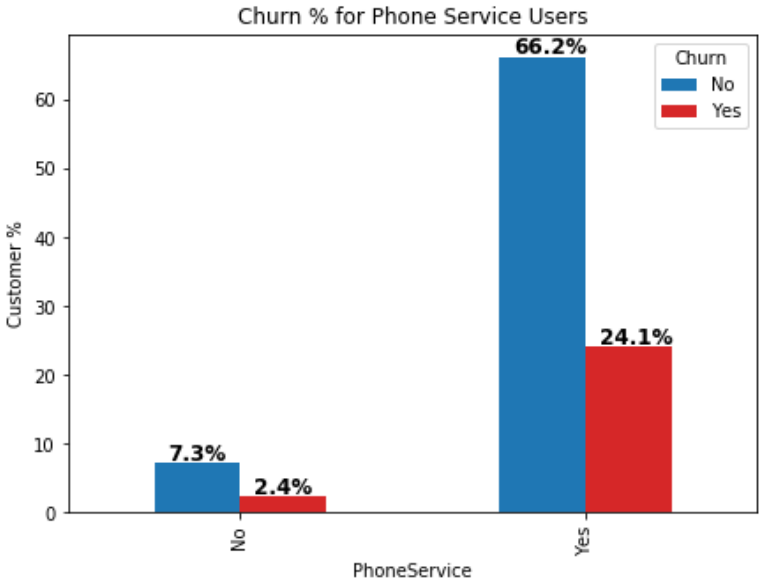


- Now we can clearly see that the churn rate is high within the 1st year.

2. **Service specific attributes -**

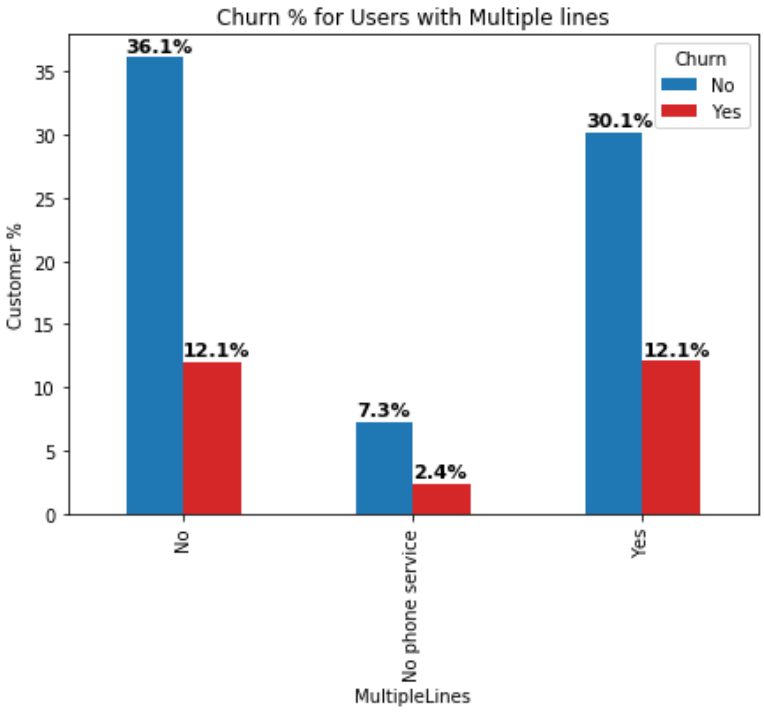
   a. Phone:

   Phone Service



- Customers having phone service have higher churn rate.

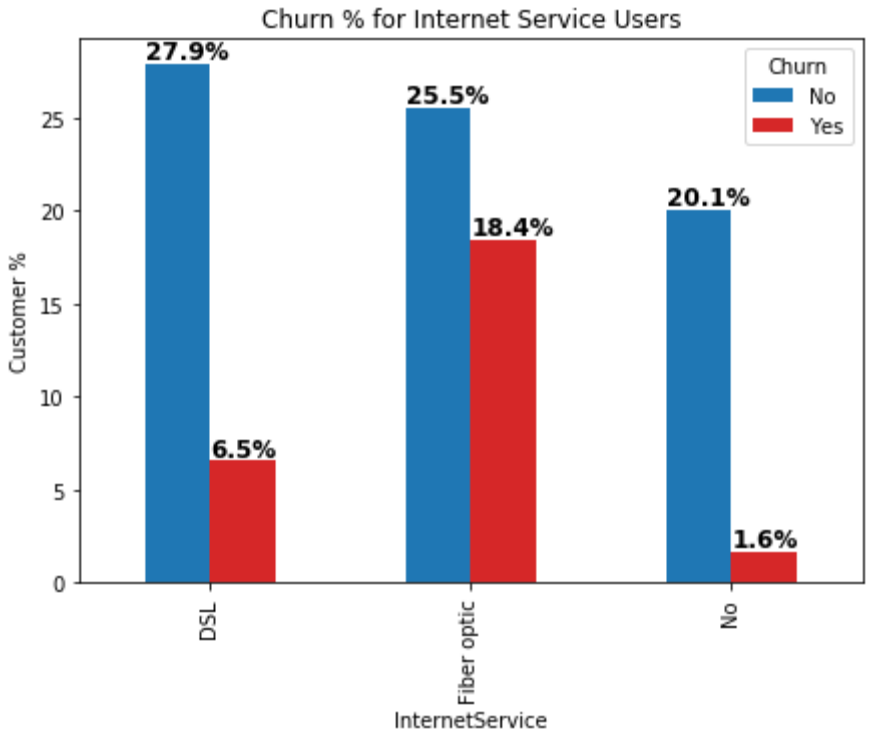
  Multiple Lines



- Customers having multiple lines or not does not affect the churn rate.
- Customers without phone service tend to churn less.

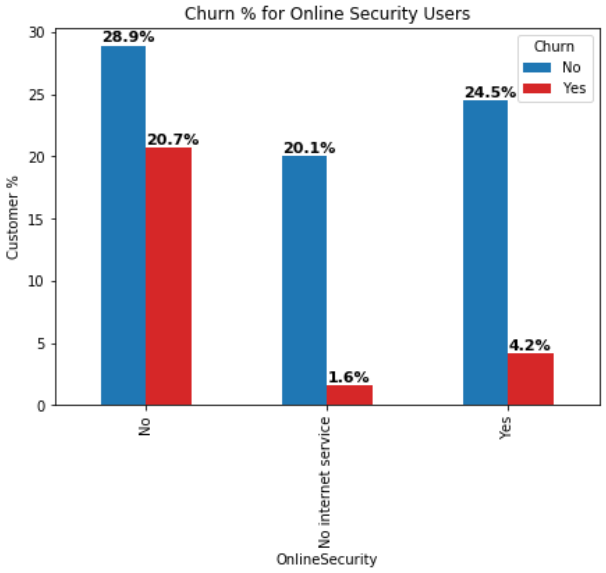
b. Internet:

Internet Service

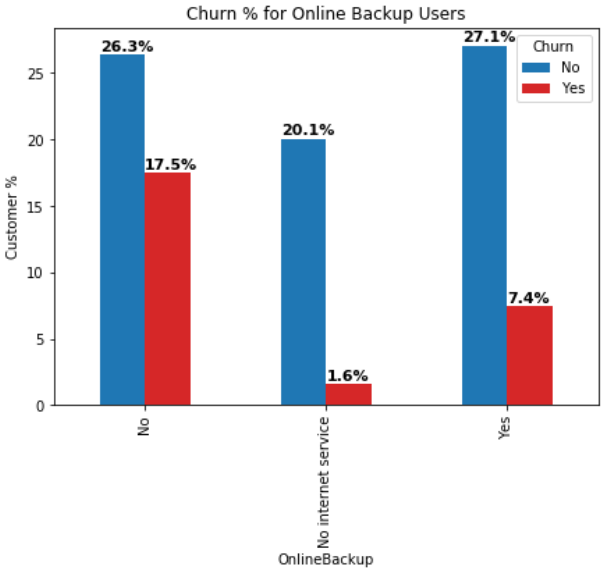


- Customers with fiber optic connection have higher churn rate.

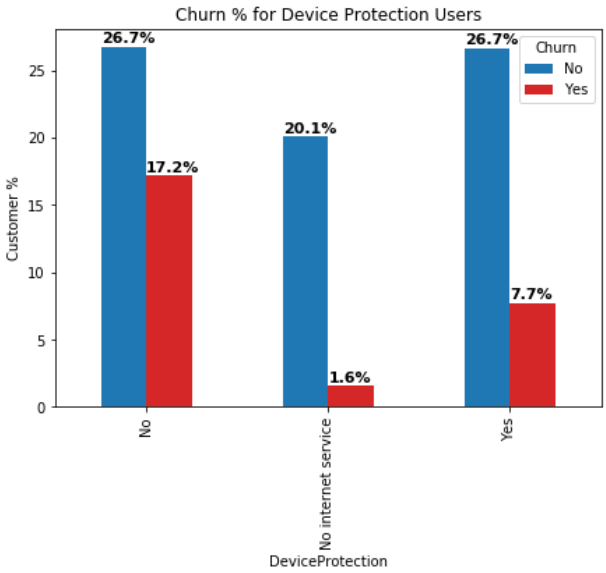
Online Security



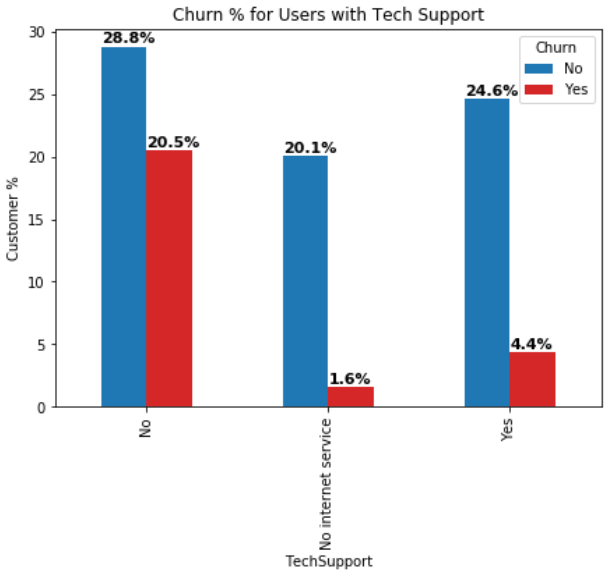
Online Backup



Device Protection

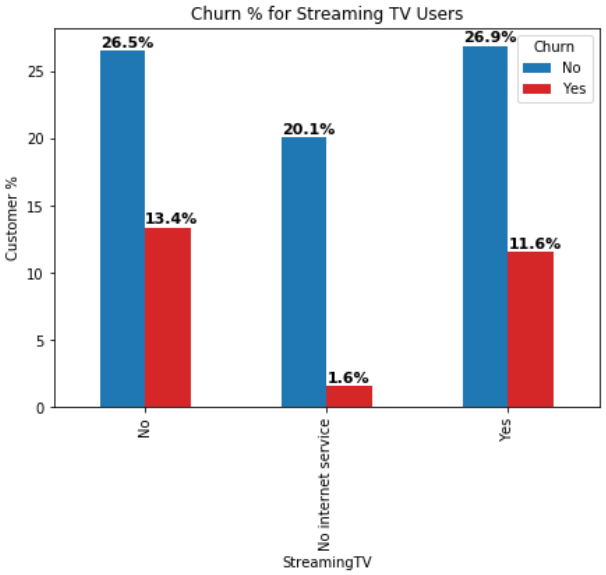


Tech Support

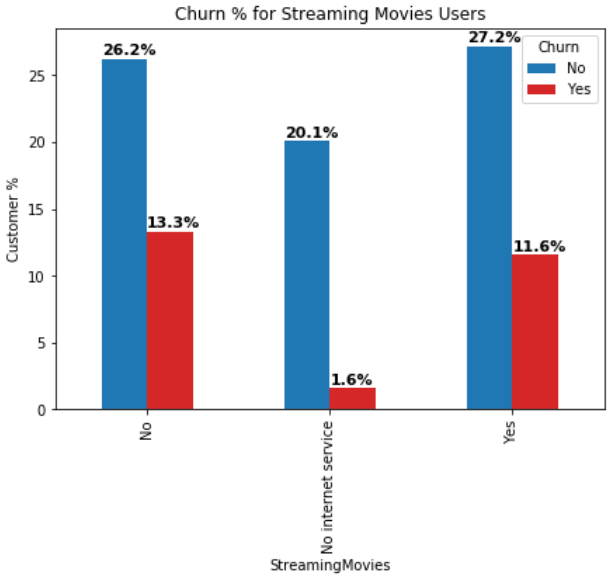


- Customers who do not have Online Security, Online Backup, Device Protection and Tech Support have higher churn rate.

Streaming TV



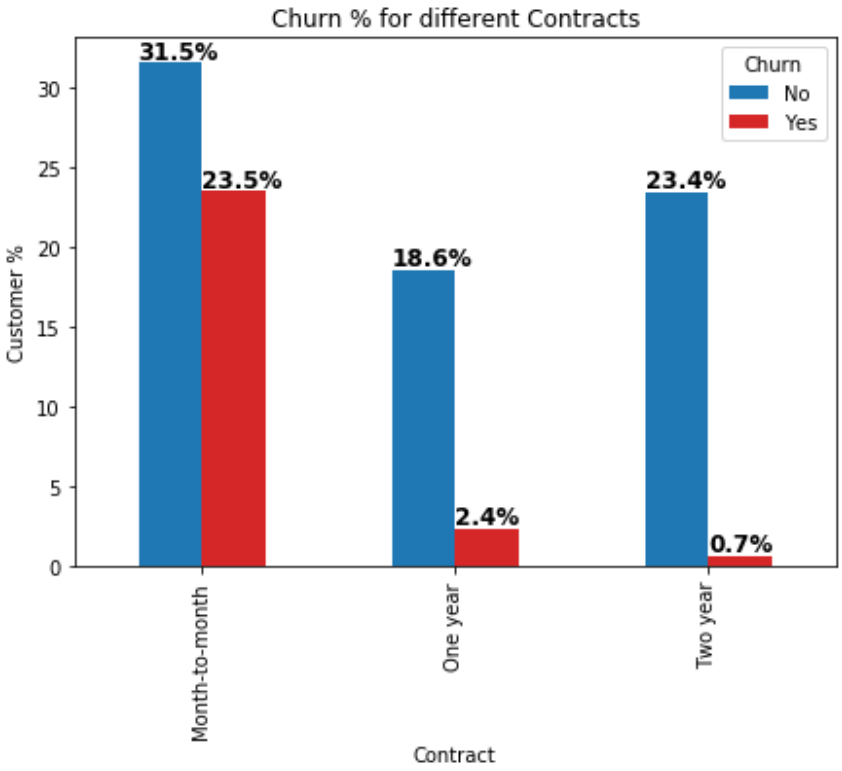
Streaming Movies



- The churn rate do not have a big difference between the customers having the service of Streaming TV & Streaming Movies or not.

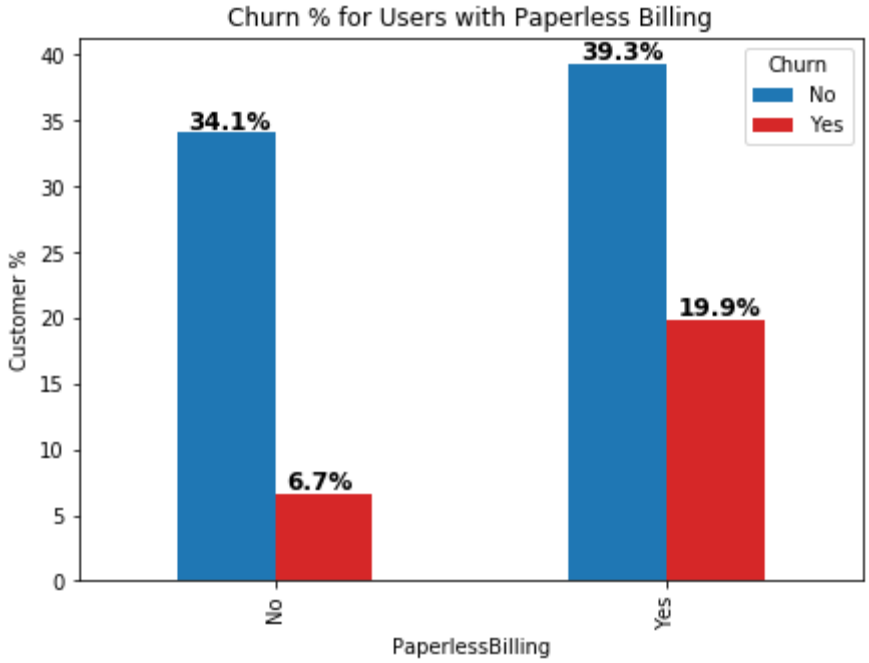
3. Money specific attributes -

Contract



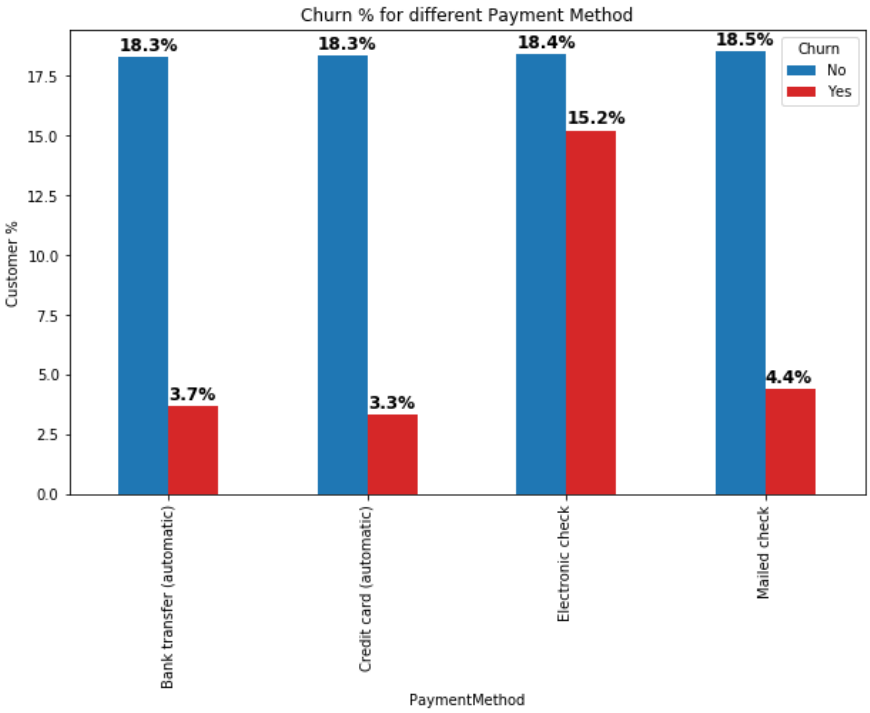
- Customers having Month-to-Month contracts have a high churn rate.

Paperless Billing



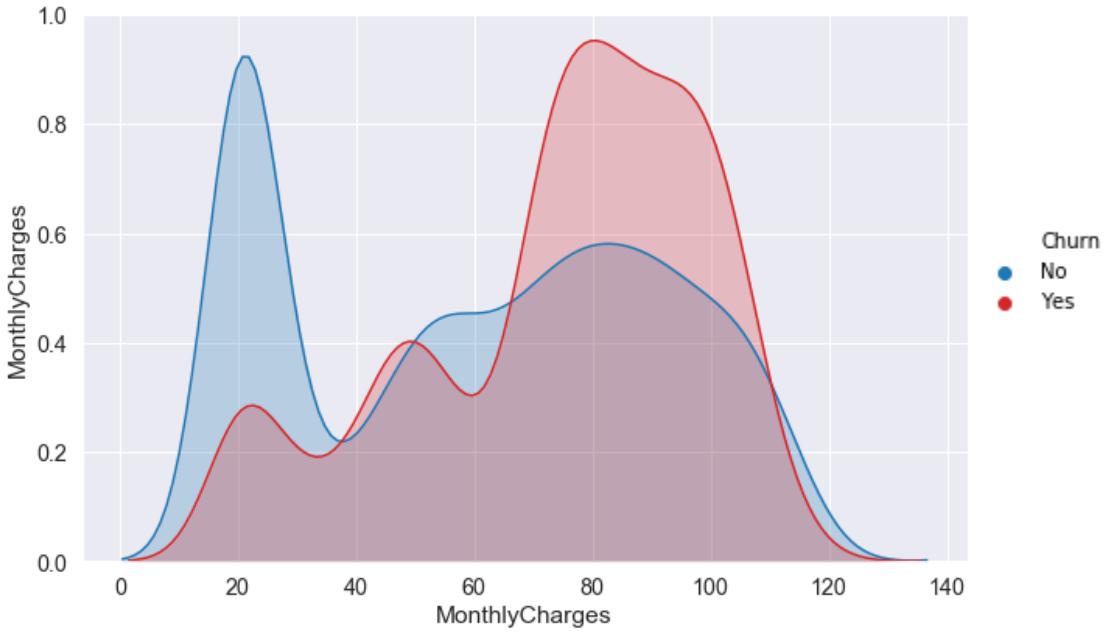
- Customers with paperless billing tend to churn out more.

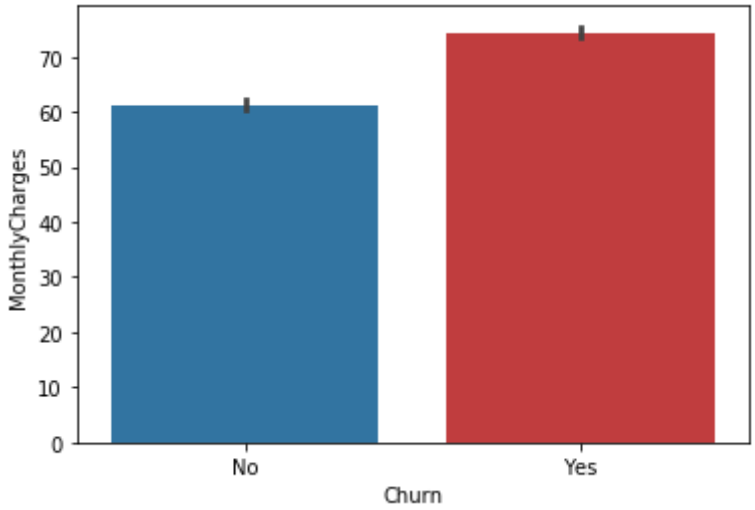
Payment Method



- Customers who pay electronic check have a high churn rate.

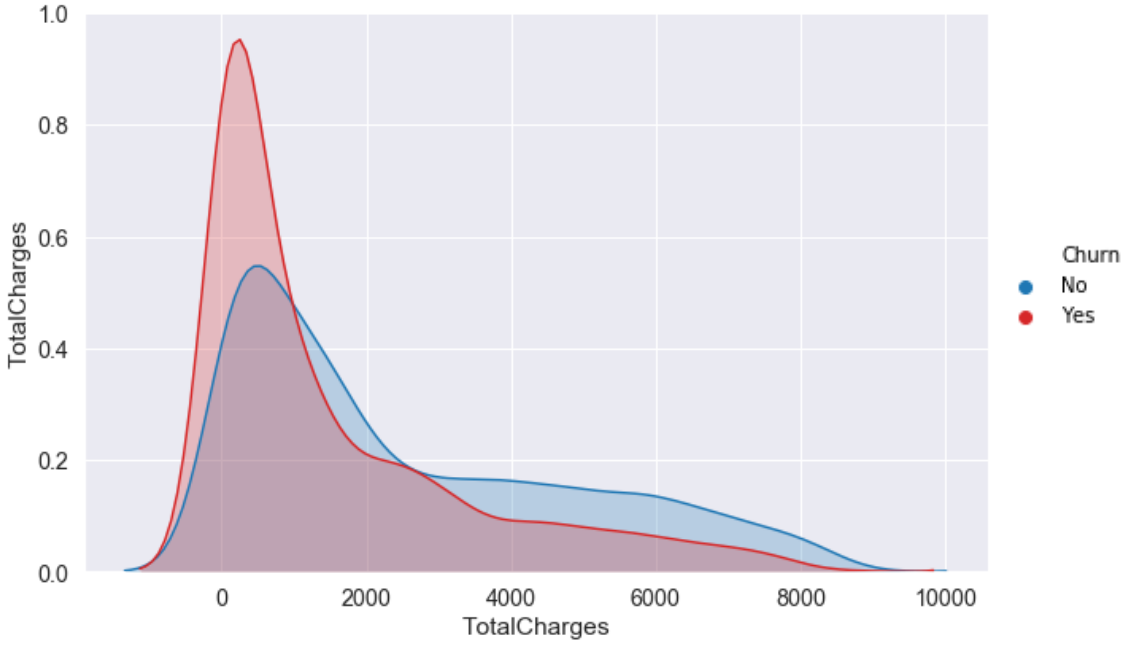
Monthly Charges

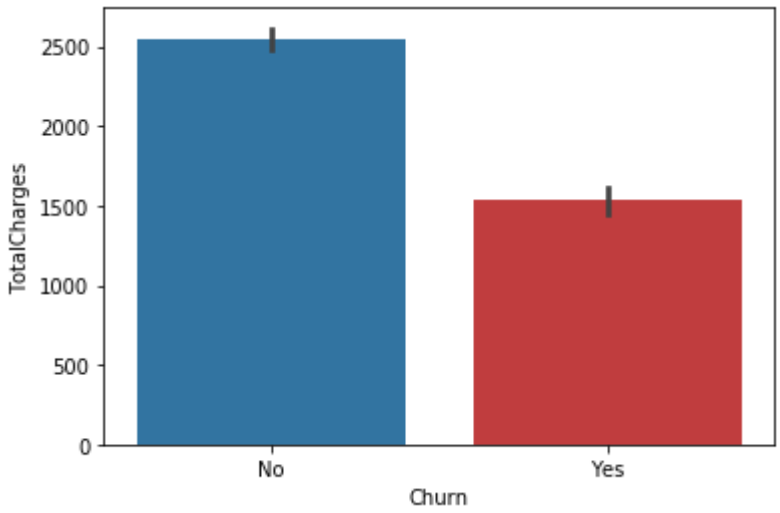




- Churn rate increases as Monthly Charges increases.

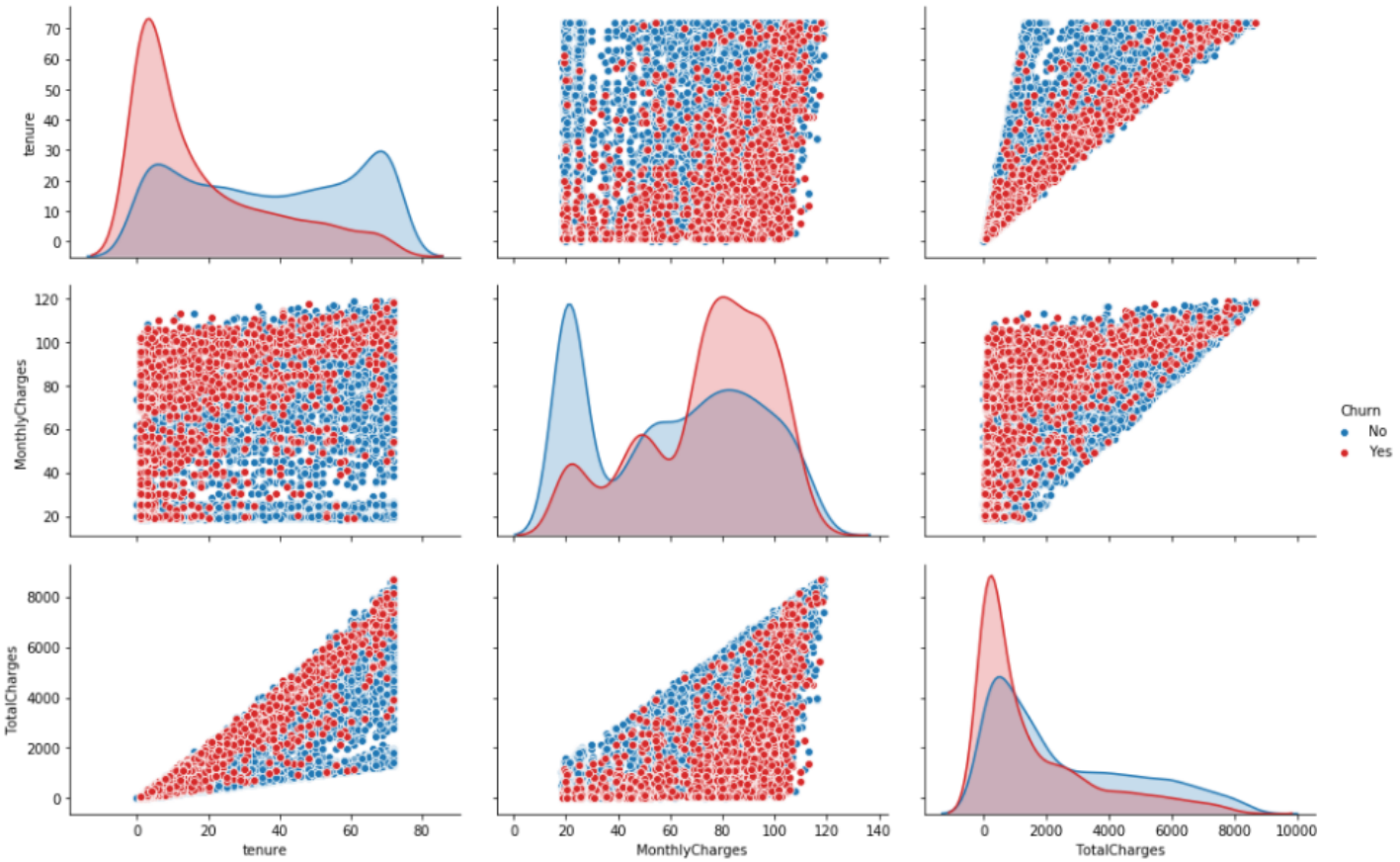
Total Charges



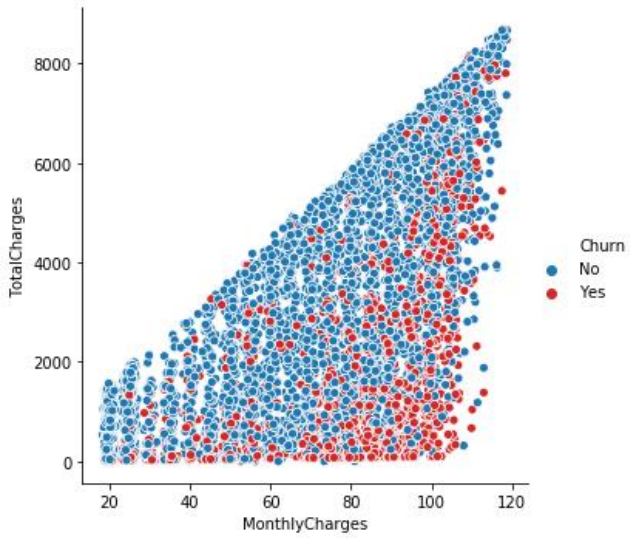


- Churn rate decreases as Total Charges increases.

Relationship between the numeric columns with respect to Churn:

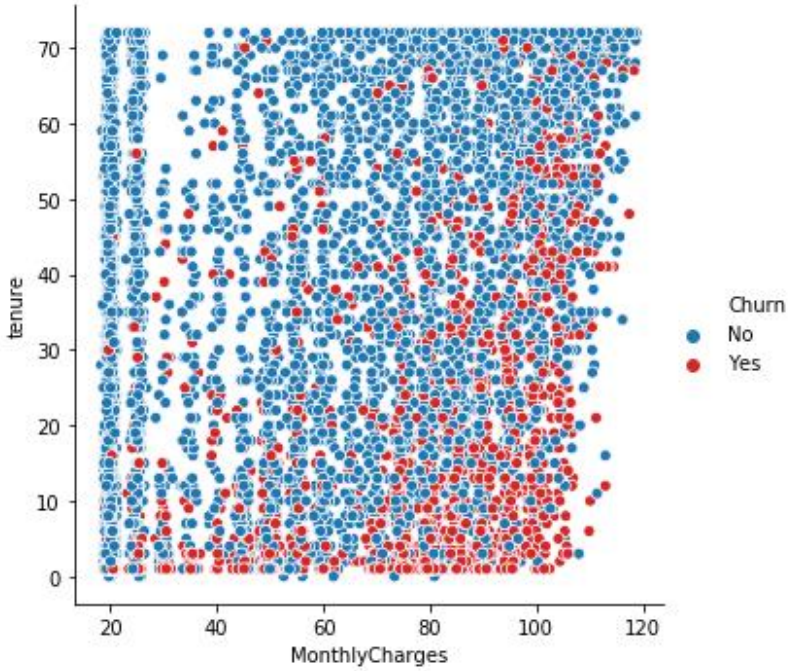


Monthly Charges and Total Charges by Churn -



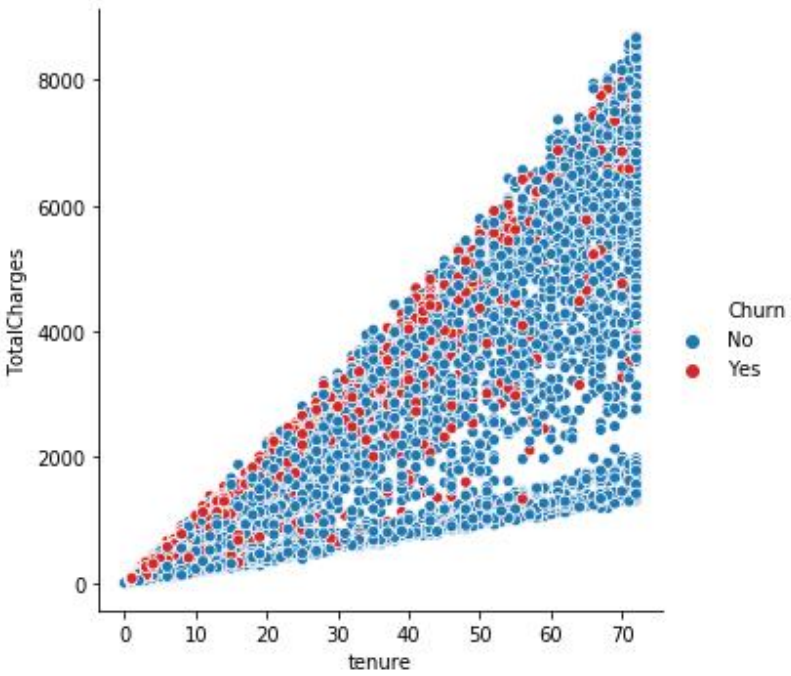
- Total Charge increases as the Monthly Charge increases.
- Churn is mainly towards the bottom which indicates that churn increases with increase in monthly charge.

Monthly Charges and Tenure by Churn -



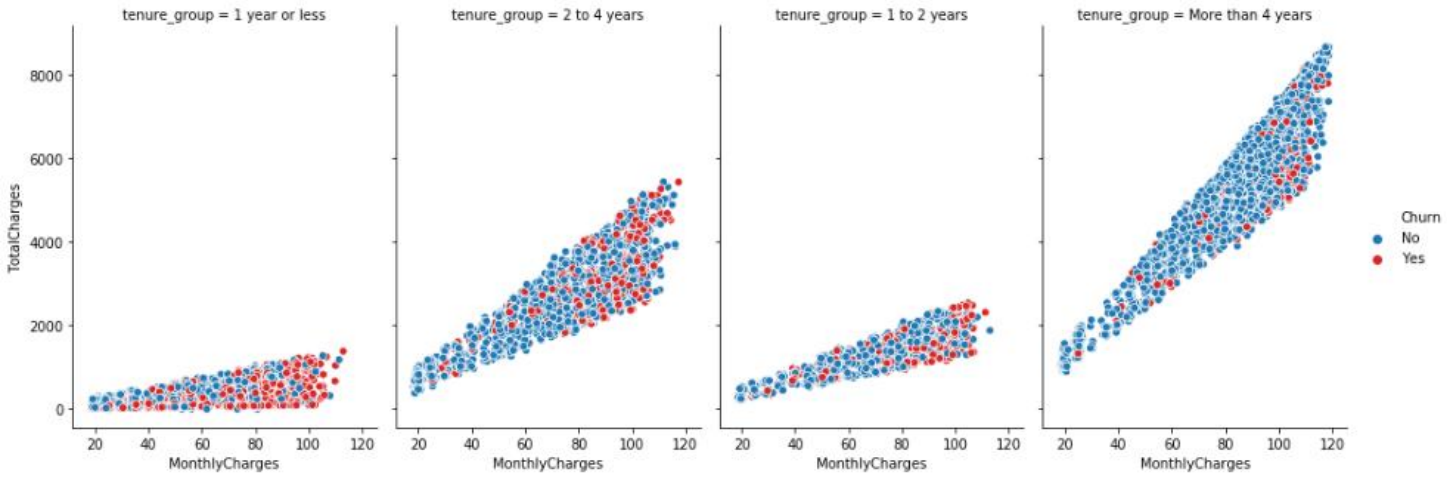
- Monthly charges may or may not increase with tenure.
- Again we can see that churn increases with increase in monthly charge.

Total Charges and Tenure by Churn -



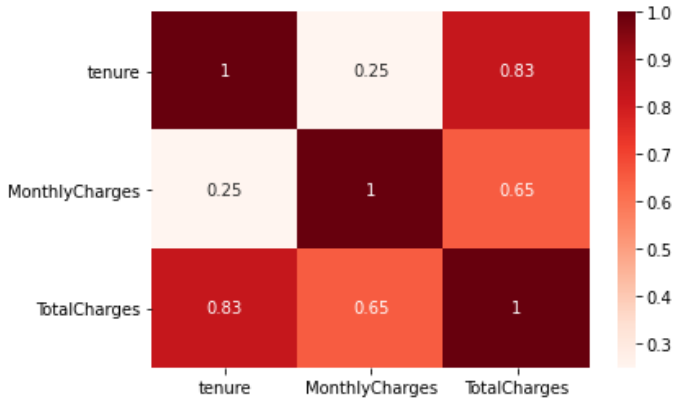
- Total charges increases with tenure.
- Churn rate does not increase so much with increase of either total charge or tenure.

Monthly Charges, Total Charges and Churn by Tenure groups -



- This clearly indicates that the churn rate is high within the 1st year and also it increases with monthly charge.

Correlation between the numeric columns:



- Tenure is highly correlated with total charge.
- Monthly charge is moderately correlated with total charge.

Inferential Statistics

Here we will be applying statistical tools to gain some inferences and insights into the data and discover relationships between various features of our dataset with the target variable by hypothesis testing.

We will use chi-square test of independence of variables in a contingency table.

Gender influence on Churn:

Let us check if gender has any influence on churn or not.

- **Null Hypothesis:** Gender has no influence on churn.
- **Alternate Hypothesis:** Gender has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 0.4840828822091383
- **p-value:** 0.48657873605618596

The p-value obtained is more than our chosen significance level. Therefore, we accept the null hypothesis, i.e. gender has no influence on churn.

Senior Citizen influence on Churn:

Let us check if being a senior citizen has any influence on churn or not.

- **Null Hypothesis:** Senior citizen has no influence on churn.
- **Alternate Hypothesis:** Senior citizen has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 159.42630036838742
- **p-value:** 1.51×10^{-36}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. senior citizen has influence on churn.

Partner influence on Churn:

Let us check if having a partner has any influence on churn or not.

- **Null Hypothesis:** Partner has no influence on churn.

- **Alternate Hypothesis:** Partner has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 158.7333820309922
- **p-value:** 2.14×10^{-36}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. partner has influence on churn.

Dependent influence on Churn:

Let us check if having a dependent has any influence on churn or not.

- **Null Hypothesis:** Dependents has no influence on churn.
- **Alternate Hypothesis:** Dependents has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 189.12924940423474
- **p-value:** 4.92×10^{-43}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. dependents has influence on churn.

Tenure influence on Churn:

Let us check if tenure has any influence on churn or not.

- **Null Hypothesis:** Tenure has no influence on churn.
- **Alternate Hypothesis:** Tenure has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 1065.3308567510544
- **p-value:** 1.28×10^{-176}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. tenure has influence on churn.

Phone Service influence on Churn:

Let us check if having phone service has any influence on churn or not.

- **Null Hypothesis:** Phone service has no influence on churn.
- **Alternate Hypothesis:** Phone service has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 0.9150329892546948
- **p-value:** 0.3387825358066928

The p-value obtained is more than our chosen significance level. Therefore, we accept the null hypothesis, i.e. phone service has no influence on churn.

Multiple Lines influence on Churn:

Let us check if having multiple phone lines has any influence on churn or not.

- **Null Hypothesis:** Multiple lines has no influence on churn.
- **Alternate Hypothesis:** Multiple lines has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 11.33044148319756
- **p-value:** 0.0034643829548773

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. multiple lines has influence on churn.

Internet Service influence on Churn:

Let us check if having internet service has any influence on churn or not.

- **Null Hypothesis:** Internet service has no influence on churn.
- **Alternate Hypothesis:** Internet service has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 732.309589667794
- **p-value:** 9.57×10^{-160}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. internet service has influence on churn.

Online Security influence on Churn:

Let us check if having online security has any influence on churn or not.

- **Null Hypothesis:** Online security has no influence on churn.
- **Alternate Hypothesis:** Online security has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 849.9989679615965
- **p-value:** 2.66×10^{-185}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. online security has influence on churn.

Online Backup influence on Churn:

Let us check if having online backup has any influence on churn or not.

- **Null Hypothesis:** Online backup has no influence on churn.
- **Alternate Hypothesis:** Online backup has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 601.812790113409
- **p-value:** 2.08×10^{-131}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. online backup has influence on churn.

Device Protection influence on Churn:

Let us check if having device protection has any influence on churn or not.

- **Null Hypothesis:** Device protection has no influence on churn.
- **Alternate Hypothesis:** Device protection has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 558.419369407389
- **p-value:** 5.5×10^{-122}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. device protection has influence on churn.

Tech Support influence on Churn:

Let us check if having tech support has any influence on churn or not.

- **Null Hypothesis:** Tech support has no influence on churn.
- **Alternate Hypothesis:** Tech support has influence on churn.

We will assume our significance level, $\boldsymbol{\alpha = 0.05}$.

Chi-square test:

- **Chi-squared test statistic:** 828.1970684587394
- **p-value:** 1.44×10^{-180}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. tech support has influence on churn.

Streaming TV influence on Churn:

Let us check if having streaming TV service has any influence on churn or not.

- **Null Hypothesis:** Streaming TV has no influence on churn.
- **Alternate Hypothesis:** Streaming TV has influence on churn.

We will assume our significance level, $\boldsymbol{\alpha = 0.05}$.

Chi-square test:

- **Chi-squared test statistic:** 374.2039433109813
- **p-value:** 5.53×10^{-82}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. streaming TV has influence on churn.

Streaming Movies influence on Churn:

Let us check if having streaming movies service has any influence on churn or not.

- **Null Hypothesis:** Streaming movies has no influence on churn.
- **Alternate Hypothesis:** Streaming movies has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 375.6614793452656
- **p-value:** 2.66×10^{-82}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. streaming movies has influence on churn.

Contract influence on Churn:

Let us check if a type of contract has any influence on churn or not.

- **Null Hypothesis:** Contract has no influence on churn.
- **Alternate Hypothesis:** Contract has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 1184.5965720837926
- **p-value:** 5.86×10^{-258}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. contract has influence on churn.

Paperless Billing influence on Churn:

Let us check if paperless billing has any influence on churn or not.

- **Null Hypothesis:** Paperless billing has no influence on churn.
- **Alternate Hypothesis:** Paperless billing has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 258.27764906707307
- **p-value:** 4.07×10^{-58}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. paperless billing has influence on churn.

Payment Method influence on Churn:

Let us check if payment method has any influence on churn or not.

- **Null Hypothesis:** Payment method has no influence on churn.
- **Alternate Hypothesis:** Payment method has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 648.1423274814
- **p-value:** 3.68×10^{-140}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. payment method has influence on churn.

Monthly Charge influence on Churn:

Let us check if monthly charge has any influence on churn or not.

- **Null Hypothesis:** Monthly charge has no influence on churn.
- **Alternate Hypothesis:** Monthly charge has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 2123.609129997958
- **p-value:** 1.88×10^{-18}

The p-value obtained is less than our chosen significance level. Therefore, we reject the null hypothesis and accept the alternate hypothesis, i.e. monthly charge has influence on churn.

Total Charge influence on Churn:

Let us check if total charge has any influence on churn or not.

- **Null Hypothesis:** Total charge has no influence on churn.
- **Alternate Hypothesis:** Total charge has influence on churn.

We will assume our significance level, α = **0.05**.

Chi-square test:

- **Chi-squared test statistic:** 6514.047812769442
- **p-value:** 0.5532461954861401

The p-value obtained is more than our chosen significance level. Therefore, we accept the null hypothesis, i.e. total charge has no influence on churn.

The chi-square test showed us that Gender, Phone Service & Total Charges have no influence on Churn. 16 out of the 19 variables have influence on churn.

Machine Learning

Data Pre-processing

As we move from EDA to machine learning, some pre-processing needs to be done in order to build our models.

We have 18 categorical and 3 numerical features. Since scikit-learn does not accept categorical features, we need to encode them. We do this by 'Label Encoding' the categorical features. In the next step we standardize the 3 numerical columns.

Now we look at the correlation matrix again to see any new correlation since all the columns are numeric now after label encoding. We can see the high correlation of tenure - total charges and monthly charges - total charges, same like we saw in EDA.

We now visualize the principal components using the principal component analysis. The 2 principal components represent atleast 50% variance in the data. We plot these 2 principal components.

Model Building

First we define some functions to tune the parameters and see the performance of the model. Then we split the data into training and test sets on a 80:20 ratio.

Baseline Model:

The next step we build a baseline model. Here we see the performance of the models using default parameters. The classifiers we choose to train and test our data are:

1. Logistic Regression
2. K-Nearest Neighbors
3. Support Vector Machine
4. Decision Tree
5. Random Forest
6. Gaussian Naive Bayes

The accuracy of the classifiers with default parameters are:

Classifier	Accuracy Score
Logistic Regression	79%
K-Nearest Neighbors	76%
Support Vector Machine	79%
Decision Tree	71%
Random Forest	77%
Gaussian Naive Bayes	73%

Logistic Regression and SVM performed the best with an accuracy of 79% and Decision Tree performed the worst with an accuracy of 71%.

Parameter Tuning:

Now we try to improve the performance of the models with parameter tuning. We will use both Grid Search CV and Randomized Search CV.

1. Logistic Regression:-

The following parameter was tuned:

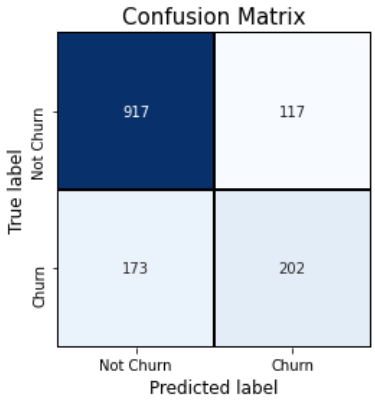
- 'C': [0.001, 0.01, 0.1, 1, 10, 100]

The best value obtained was C = 0.1.

We got a slight increase of 0.28% in accuracy.

Accuracy = 79%

Confusion matrix =



Precision = 63%

Recall = 54%

f1-score = 58%

AUC = 82%

Feature importance = Monthly charge is highly positively related with churn whereas tenure, phone service & contract are highly negatively related to churn.

2. K-Nearest Neighbors:-

The following parameter was tuned:

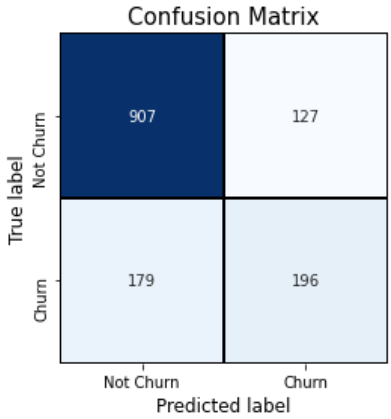
- 'n_neighbors': (1 to 49)

The best value obtained was n_neighbors = 38.

We get an increase of 2.6% in accuracy.

Accuracy = 78%

Confusion matrix =



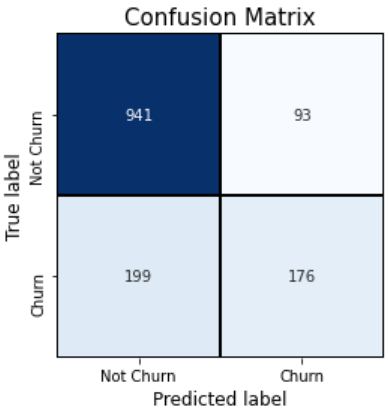
Precision = 61%
Recall = 52%
f1-score = 56%
AUC = 81%

3. Support vector Machine:-

The following parameters were tuned:

- 'C': (1 to 19)
- 'kernel': ['linear', 'poly', 'rbf', 'sigmoid']
- gamma': ['scale', 'auto']

The best values obtained were C = 1, kernel = 'rbf', gamma = 'scale'.
Accuracy remains the same.
Accuracy = 79%
Confusion matrix =



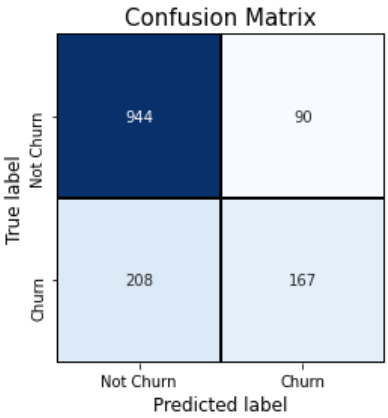
Precision = 65%
Recall = 47%
f1-score = 55%
AUC = 78%

4. Decision Tree:-

The following parameters were tuned:

- 'max_depth': (2 to 6)
- 'max_features': (1 to 20)
- 'min_samples_leaf': (1 to 19)
- 'criterion': ['gini', 'entropy']

The best values obtained were min_samples_leaf = 4, max_features = 13, max_depth = 5, criterion = 'entropy', random_state = 10.
We get a huge increase of 7.24% in accuracy.
Accuracy = 79%
Confusion matrix =



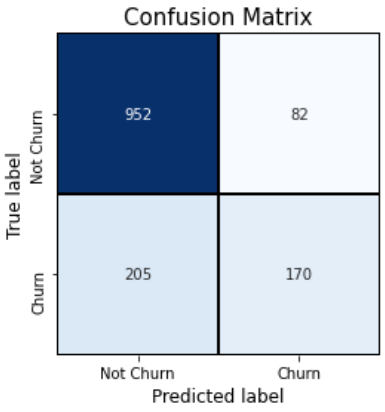
Precision = 65%
Recall = 45%
f1-score = 53%
AUC = 81%
Feature importance = We get tech support and monthly charges highly related to churn.

5. Random Forest:-

The following parameters were tuned:

- 'max_depth': (2 to 6)
- 'max_features': ['sqrt']
- 'min_samples_leaf': (1 to 19)

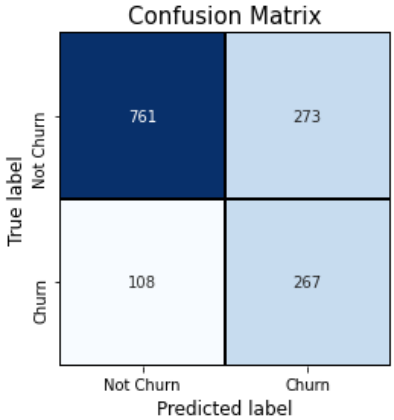
The best values obtained were min_samples_leaf = 13, max_features = 'sqrt', max_depth = 6.
We get an increase of nearly 2.4% in accuracy.
Accuracy = 79%
Confusion matrix =



Precision = 67%
Recall = 45%
f1-score = 54%
AUC = 83%
Feature importance = We get contract and tenure highly related to churn.

6. Gaussian Naive Bayes:-

No parameters were tuned here.
So the accuracy remains the same.
Accuracy = 73%
Confusion matrix =



Precision = 49%
Recall = 71%
f1-score = 58%
AUC = 80%

The performance of the classifiers after tuning the parameters are:

Classifier	Accuracy Score
Logistic Regression	79% (0.23% ↑)
K-Nearest Neighbors	78% (2.6% ↑)
Support Vector Machine	79%
Decision Tree	79% (7.24% ↑)
Random Forest	79% (2.4% ↑)
Gaussian Naive Bayes	73%

After tuning the parameters we saw some good improvements in the performance of the models. Decision Tree improved the most by an increase of 7.24% in accuracy. Logistic Regression, K-Nearest Neighbors and Random Forest improved their performances too. Support Vector Machine and Gaussian Naive Bayes remained the same. The Gaussian Naive Bayes model is the lowest performing model with an accuracy of 73%.

All the models are performing equally (except the Gaussian Naive Bayes model) with an accuracy of 79% (approx).

Voting Ensemble:

We now use the Voting Classifier to combine all the above mentioned machine learning classifiers. The models are pitted against each other and selected upon best performance by voting. Such a classifier can be useful for a set of equally well performing model in order to balance out their individual weaknesses.

The voting ensemble gives an accuracy score of 80%.

Conclusion

Recommendations to retain the customers:

Take the following actions immediately:

- Try striking a longer contract with new customers: two year or more.
- Lower the monthly charges.
- Leverage the time to improve the quality of services, on the high cost ones like fiber optic.
- Improve on the Technical support on all services like streaming, phone connection and internet. Be up-to-date with current technology.
- Collect customer feedback and act on it immediately to prevent new customer churn.

Next: It will be helpful to get more features to understand the behaviour of the customers according to location and different service providers.

Scope for future work:

- More predictive models could be tried. However, there is no guarantee of better accuracy, as we have seen similar accuracy with logistic regression, decision tree and random forest. This actually means most of the variance in the data is explained.
- One could collect more data through surveys, analyze them using NLP techniques and take more measures.
- There is a scope to collect historical data on company customers over a few decades, and fight out clear reason for customer drop.