Telecom Users Churn

Project Report

Project Overview

It is very important for any telecom organization to retain the customers ,as it would cost much higher to attract a new customers .And also it is harder to work with new customers as company would not be having past information of the customer. So predicting the churn, we can react in time and retain the customer.

In this project, I have analyzed the data to identify the relation between customer's information, demographics, customer's services usage history and churning. I have used various feature selection techniques to identify important features that influence the customer behavior and fit a machine learning model to identify the potential churn.

Problem Statement

The goal is to effectively predict the behavior of customers to retain them. The tasks involved are:

- Exploratory data analysis to draw some insights from data
- Perform feature engineering and feature selection
- If any, deal with unbalanced data set
- Chose the right performance metric and fit multiple ML models and compare their results.

Metrics

A glance at class distribution reveals data is unbalanced, that makes classification accuracy not a right metric to measure performance.

Since companies aim to retain the customers as much as possible, identification of churning is of a paramount importance and here type 1 error (wrongly marking customer as potential churn) is less detrimental compared to type 2 error (Failed to identify the churn).So recall alongside F1 score is the right metric.

F1 = TP/(TP + 0.5 *(FP+FN))

Recall = TP/(TP+FN)

TP = number of true positives

FP = number of false positives

FN = number of false negatives

Exploratory data analysis and visualisation

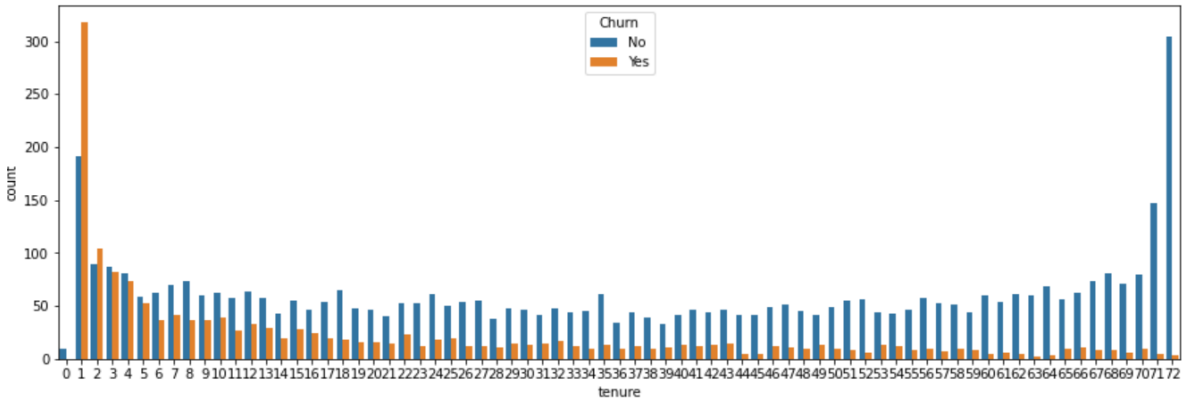
Telco user's data set consists of 5986 observations, out of which 1587 is classified as "Churn" (approximately 25%), while, others are retained.

It consists of 21 features, one of them is the target variable "Churn".

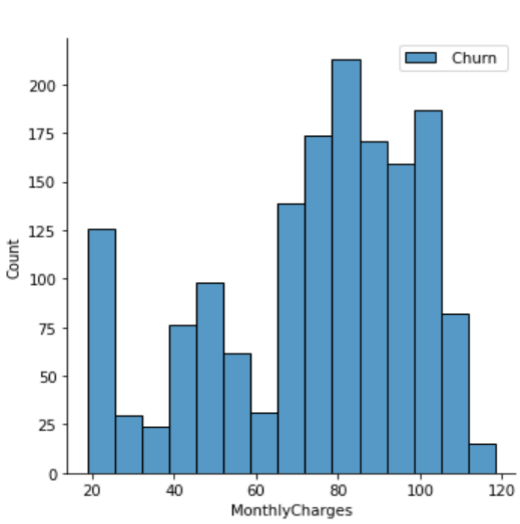
The data set do not contain much of a missing value, around 10 rows have "TotalCharges" values as empty string, corresponds to "Tenure" value 0. These values are replaced with zero as their tenure is zero.

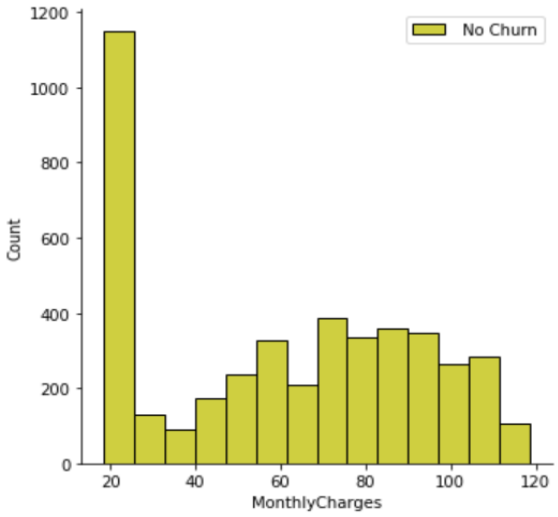
Exploratory visualization:

As can been seen in the below plot, most of the churning happening in the initials months , and as tenure goes on , it is more likely to retain the customer , particular after 65 month tenure , customer more likely to be retained.

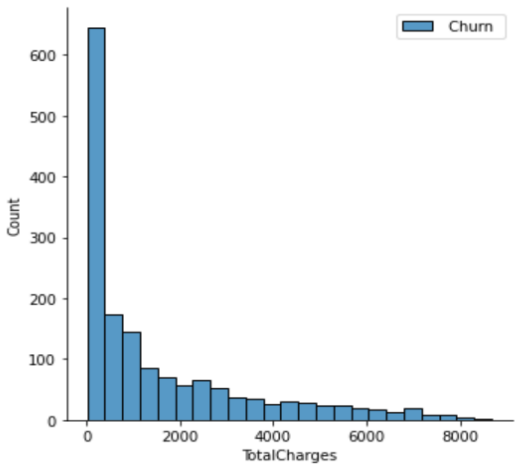


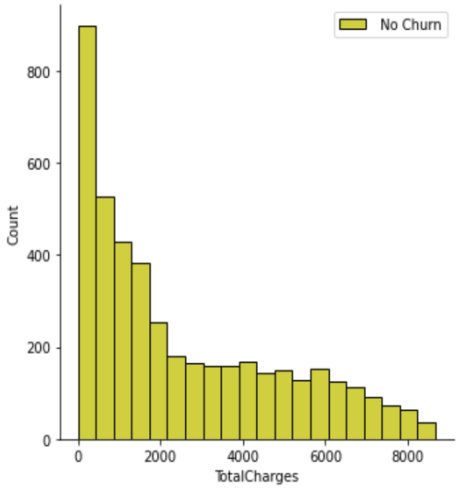
It can be observed from the below graphs, most of the churned users spend around 60-100 dollars per month, while it is much smaller for retained customers, that is 20-60. From this, we may infer, high monthly spending is one of the reasons for leaving the network.



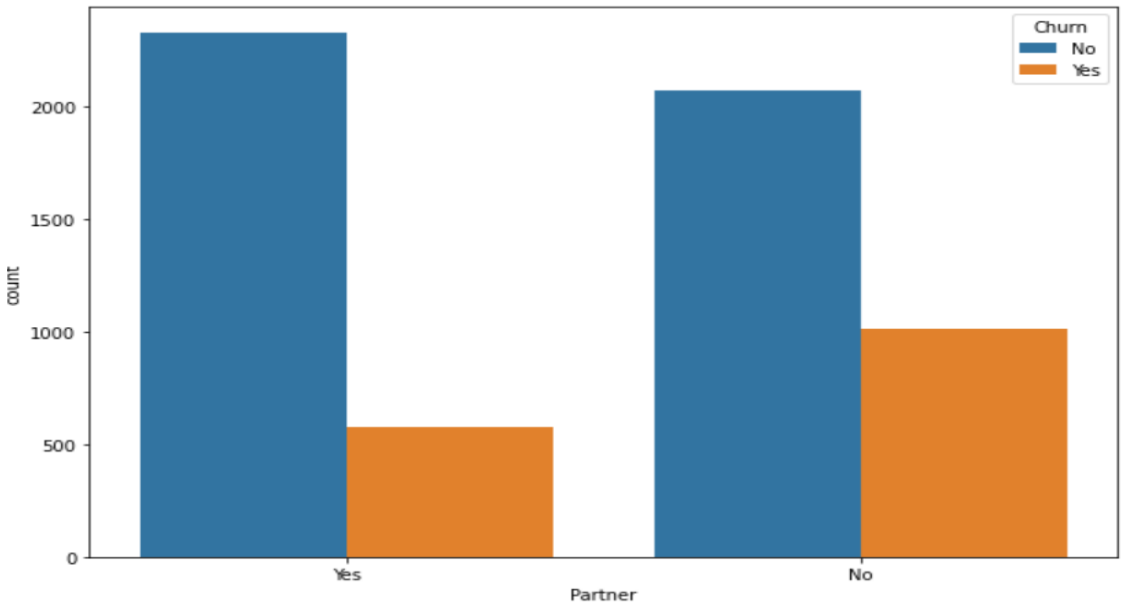


It can be seen in the below plots; number of people have high amount of 'TotalCharges' in churned data set is much less compared to retained users data set. This reiterates our earlier observation, that many customers are the leaving network in the initial months of service itself.

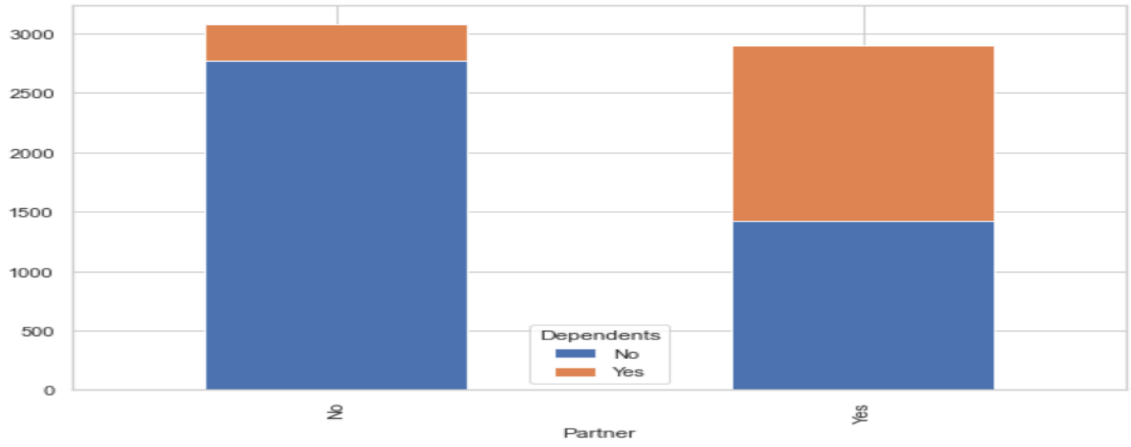




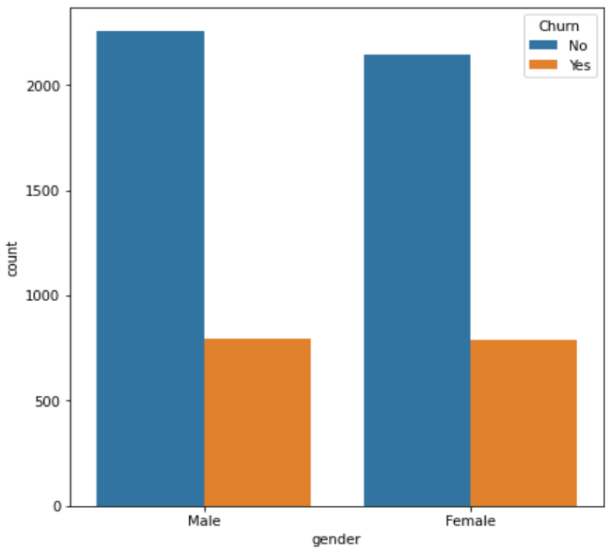
It can be seen in the below plots, people with partners are more likely to stay with the network. We can future analyze partners influence monthly charges and tenure.

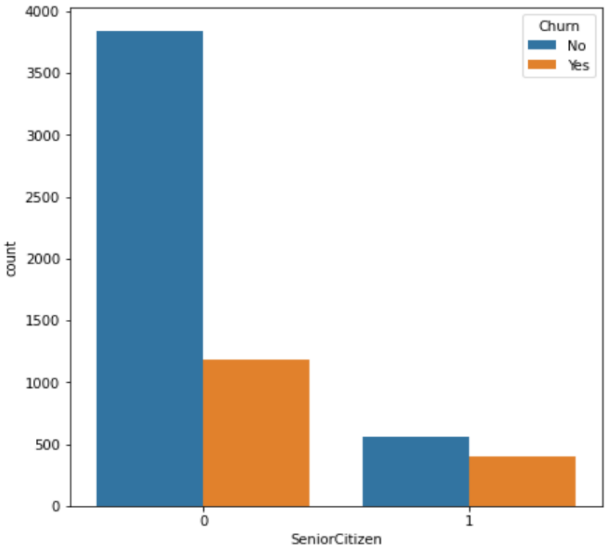


Below plots reveals, it is more likely for customers to have dependents if they have a partner,we could future analyses how dependents are influencing customers tenure and monthly charges.

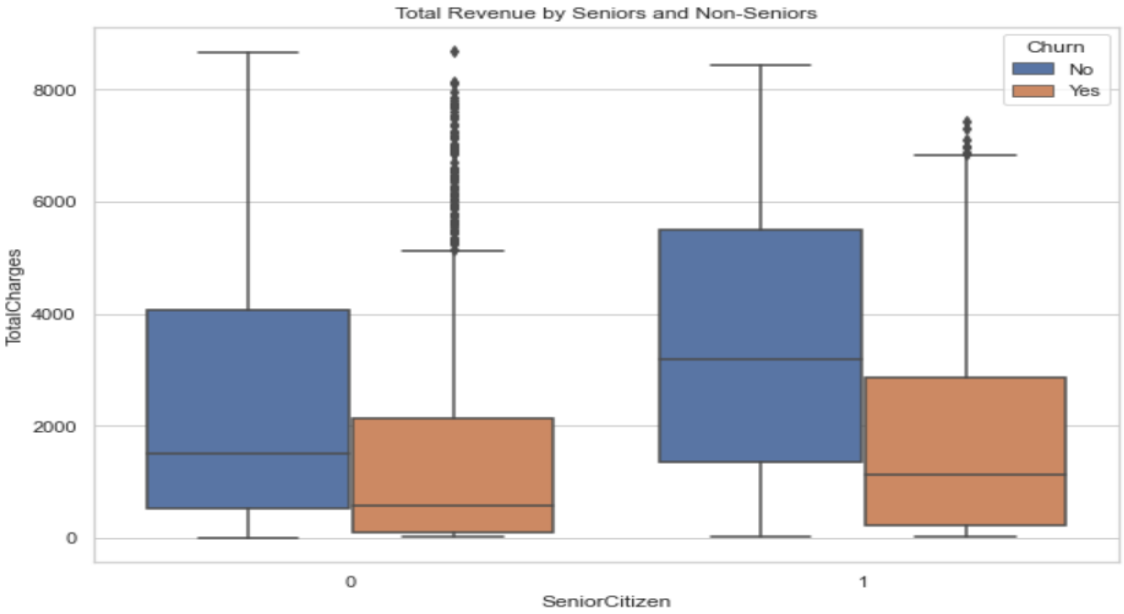


Looking into how gender and senior citizen influencing the churning, it can be observed in the below charts. Gender does not seem to be having any role while senior citizens are more likely to leave the service. We can future try analyses the factors influencing the seniors to leave the service.



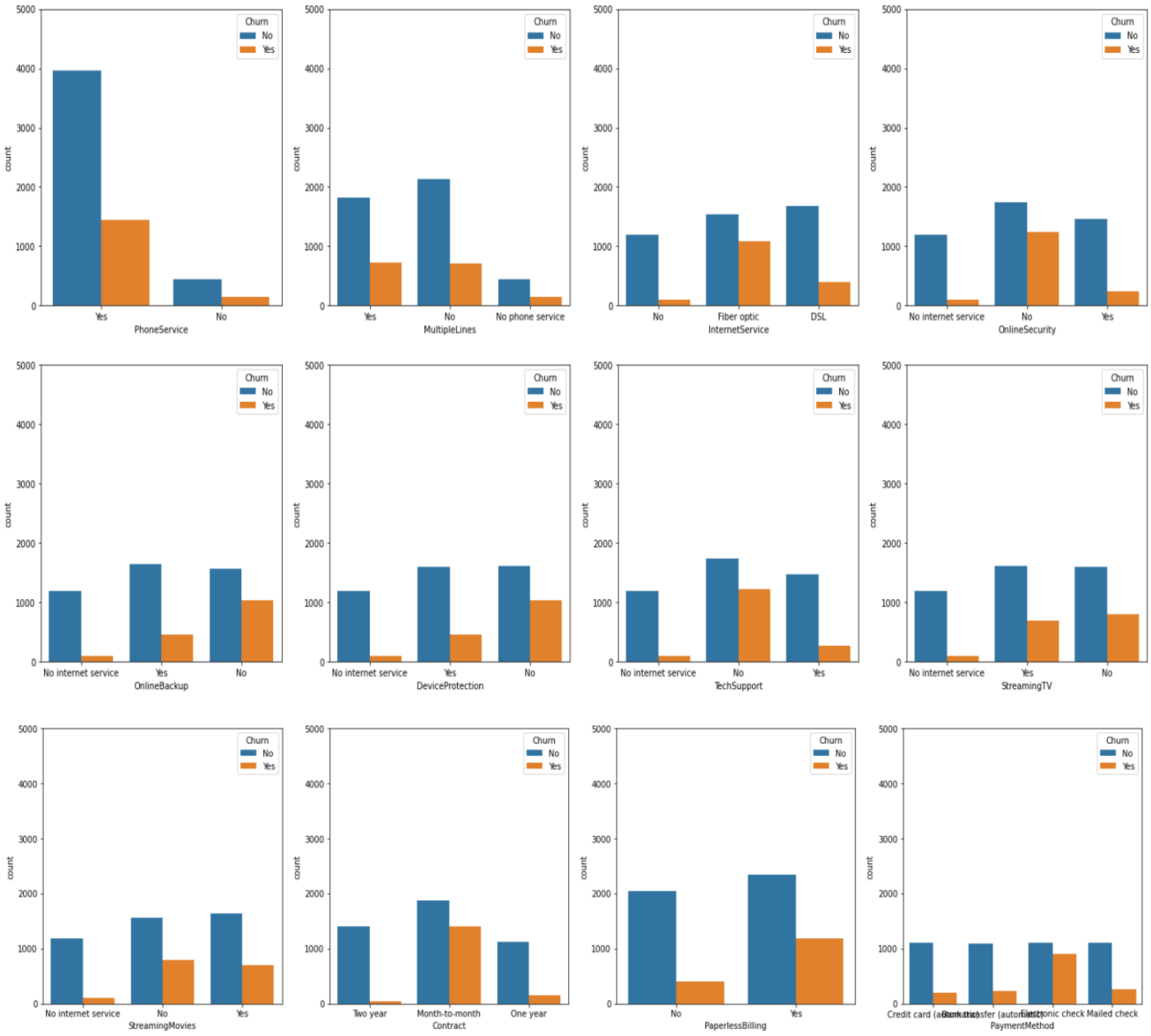


Total revenue by senior is slightly more than young customers



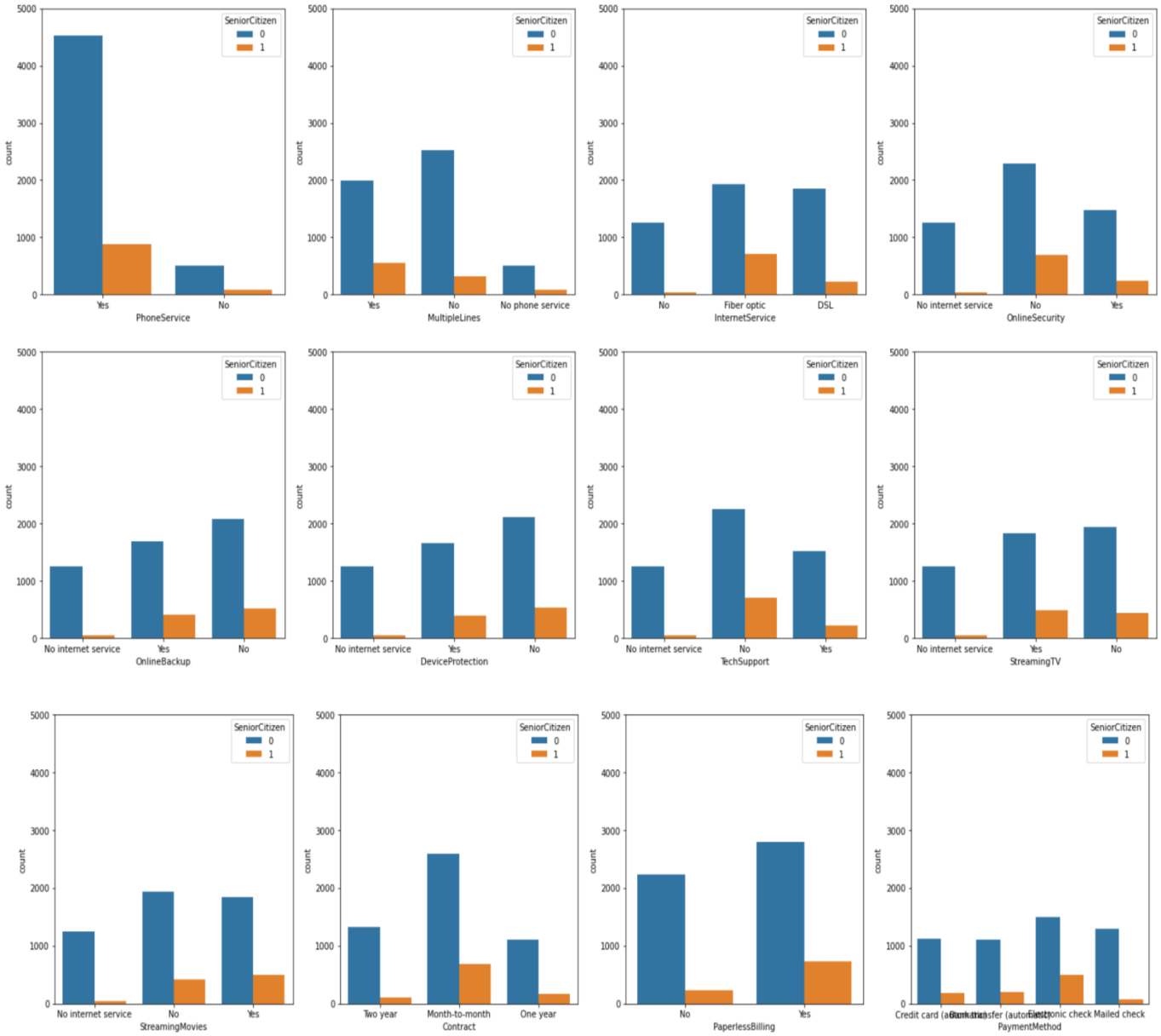
On plotting churning data against all the possible services, the observations that are more pronounced are:

- generally, people with no internet connection tend to stay with the network while people with optical fibers are more likely to leave.
- customers with internet connection but no online security, online backup, device protection or tech support is more likely to churn
- Longer the tenure, more likely for customer to stay



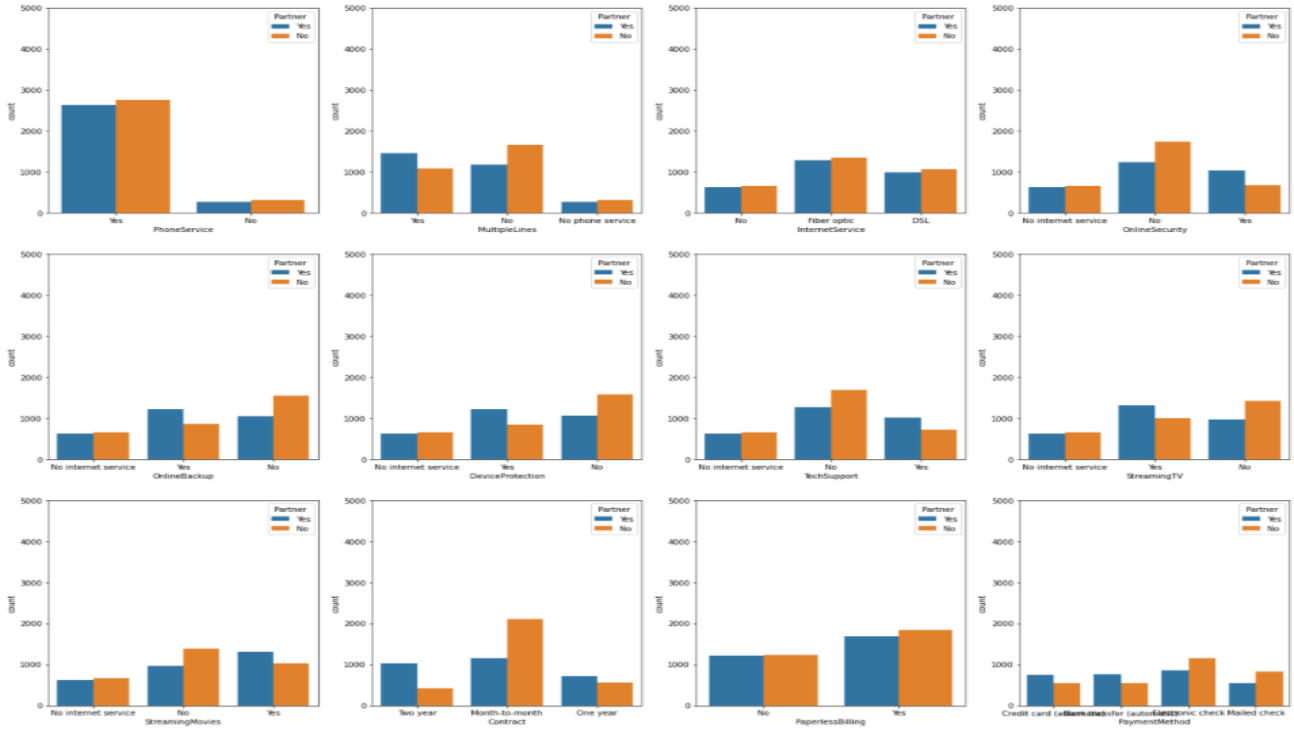
Below plots gives the distribution of services usage among young and senior people. The observations are:

- counter intuitively, there are only few senior people who do not use internet services.
- senior people more likely to use fiber optics and majority of them are on month-to-month base contract.

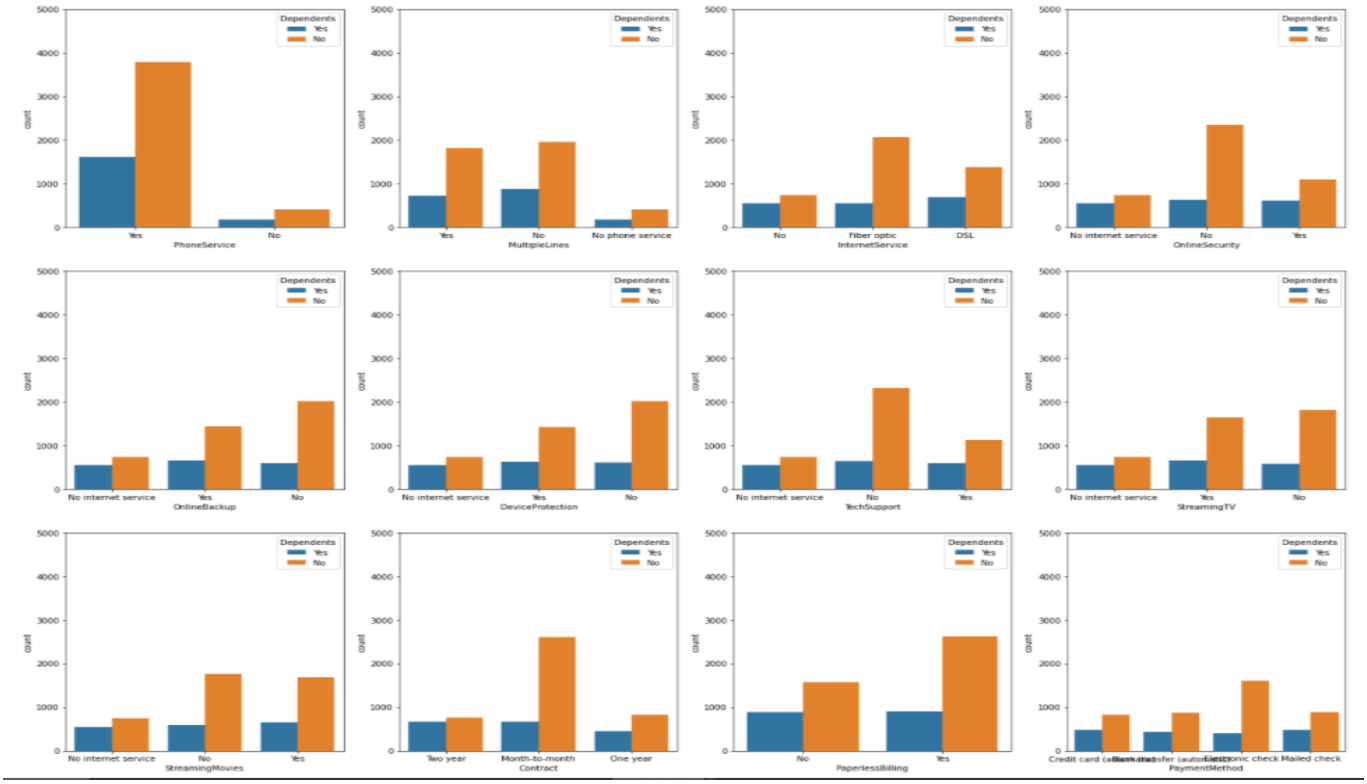


Observations regarding people who live with partner:

- They prefer month-on-month contract.



Customers with dependents prefer phoneService, Fiber optic ,online security, tech support and they are mostly on month-on-month contract.



Feature Selection

Feature selection methods are used to select most useful features to predict the target variable. It helps to reduce the complexity of the model, avoids overfitting and likely increases the performance of the model.

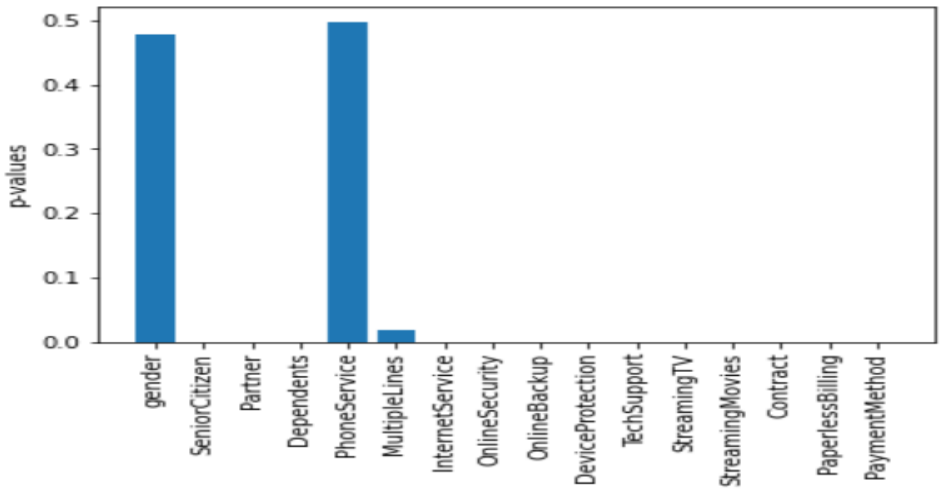
We can perform various feature selection methods based the type of the target variable and feature variable.

In our case, target variable is categorical variable, so we can perform chi-square if predictor variable is also a categorical variable. And if predictor variable is of a type numeric one-way anova is used.

Chi Square Test :

Chi-square is a test of independence between feature and target variable.

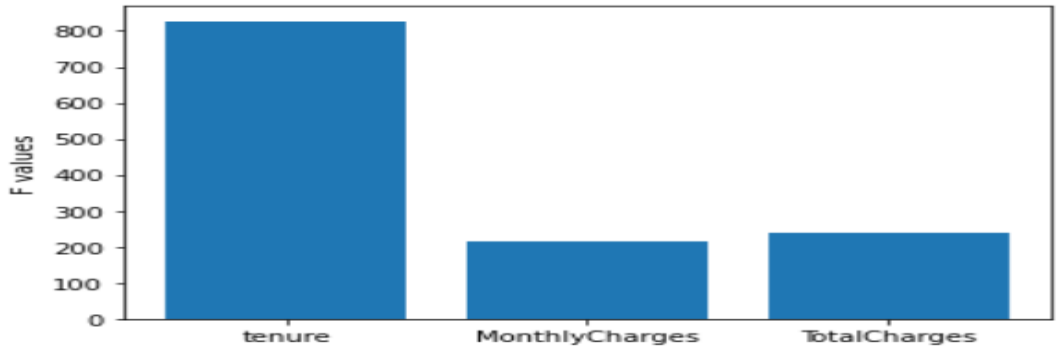
Below plot shows features 'gender' and 'PhoneService' are failing the test at significance level 5%



Anova:

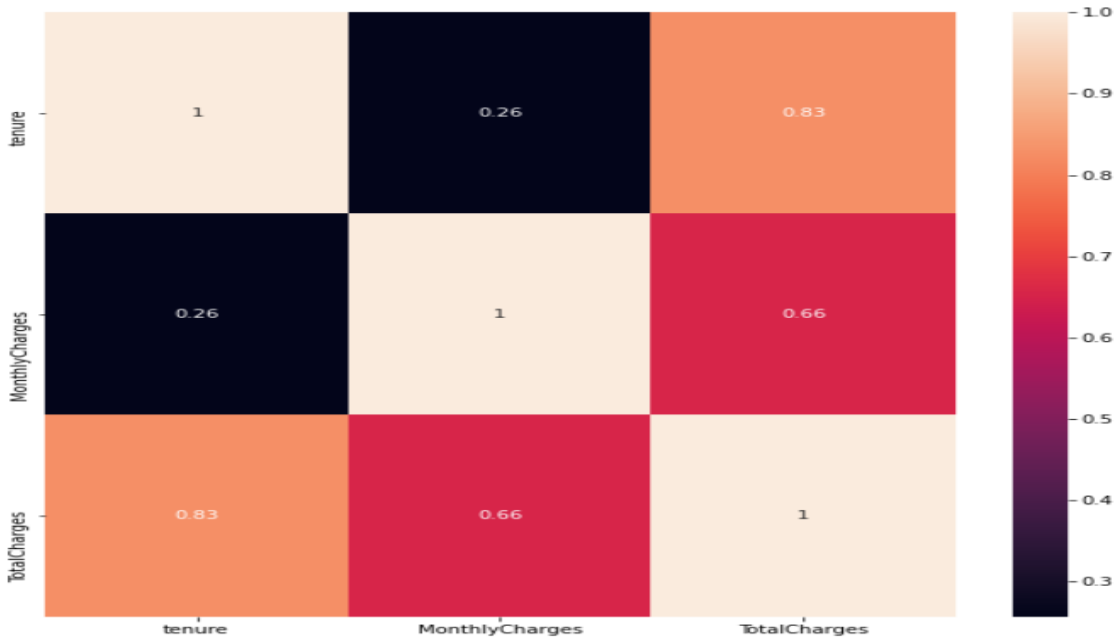
here, Target variable is categorical and few of the features are numerical ,so ,One-way anova is used to identify the feature significance of those numerical features.

Below plot shows the f values of all 3 numerical variables, which have passed significance test at 5%.



Co-relation heat map:

A heat map shows correlation coefficients between variables. A highly co-linear variables need to be avoided for better performances in many models.



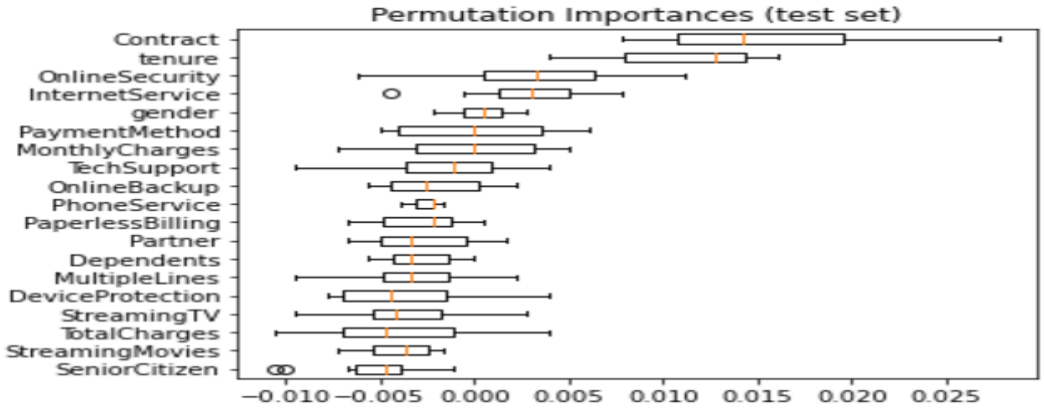
Feature importance by permutation importance:

It is the mechanism to compute the feature importance, it Shuffles the values in a single column, make predictions using the resulting dataset. Use these predictions and the true target values to calculate how much the loss function suffered from shuffling.

That performance deterioration measures the importance of the variable you just shuffled. As it can been seen in the below box plot, 'contract' ,'tenure', 'online security', 'internet

service' are the most important features while 'streaming movies' and 'seniorCitizen' are the least importance.

Based on the previous discussed models and their results, we can propose a new model, that would have only “contract”, “tenure”, “online security” and “internet services”, “payment Methods”



Model Fitting

Data is split into 70:30 ratio, Since we want to compare full model and reduced model obtained based on the observations made on feature selection technique, two separate sets of data split is done , with one them contains only few features (“contract”, “tenure”, “online security” and “internet services”,”PaymentMethod”), while , the other contains all features.

As discussed earlier, data set is unbalanced, so one of the ways of solving this problem is oversampling techniques.

Oversampling technique being used is synthetic minority oversampling technique (SMOTE), new examples are synthesized for minority class.

Here, resampling is done only on training data, not on validation or test data, as we need keep test data as close to real world situation as possible. And I am using recall and F1 score as a metric for the reasons discussed earlier.

Here, data is fit into Random Forest and Gradient Boosting algorithms. And their cross-validation scores and best parameters identified using grid searching is shown below

Random Forest:

```
CV score of random forest is: 0.7509071208190028

{'randomforestclassifier__max_depth': 5,
 'randomforestclassifier__min_samples_leaf': 1,
 'randomforestclassifier__min_samples_split': 5,
 'randomforestclassifier__n_estimators': 100}
```

Gradient Boosting:

```
CV score of Gradient boosting  is: 0.7983574071639903

{'gradientboostingclassifier__learning_rate': 0.03,
 'gradientboostingclassifier__max_depth': 2,
 'gradientboostingclassifier__min_samples_leaf': 1,
 'gradientboostingclassifier__min_samples_split': 3,
 'gradientboostingclassifier__n_estimators': 50}
```

As cross validation score of gradient boosting is higher, we can select that model to predict the test set data and calculate recall and F1 score. As can be seen below, Full model is performing significantly better than reduced model, as opposed to what was assumed based on the feature selection techniques result. I need to future investigate this phenomenon as to why my assumptions are not satisfied.

Gradient boosting:	Full model	Reduced model
recall_score	0.8340425531914893	0.46808510638297873
f_score	0.7205971759781862	0.5730796390707374

Future Improvements:

- I need to Need to future investigate on why assumptions made based on feature importance techniques are not satisfied. And identify the optimal subset of important features that would increase the performance of the model and reduces the complexity of model for better interpretability.
- Perform a relevant statistical test to verify if full model is not significantly better than reduced (subset of features) model.

Conclusion

Data is analysed to extract important insight from the data. Gradient boosting model is fit to the data and around 84% of churning can be successfully predicted.