**Project: Telecom Customer Churn Prediction**

**Introduction and Problem:**

Customer attrition, also known as customer churn, customer turnover, or customer defection, is the loss of clients or customers.

Telephone service companies, Internet service providers, pay TV companies, insurance firms, and alarm monitoring services, often use customer attrition analysis and customer attrition rates as one of their key business metrics because the cost of retaining an existing customer is far less than acquiring a new one. Companies from these sectors often have customer service branches which attempt to win back defecting clients, because recovered long-term customers can be worth much more to a company than newly recruited clients.

Companies usually make a distinction between voluntary churn and involuntary churn. Voluntary churn occurs due to a decision by the customer to switch to another company or service provider, involuntary churn occurs due to circumstances such as a customer's relocation to a long-term care facility, death, or the relocation to a distant location. In most applications, involuntary reasons for churn are excluded from the analytical models. Analysts tend to concentrate on voluntary churn, because it typically occurs due to factors of the company-customer relationship which companies control, such as how billing interactions are handled or how after-sales help is provided.

Predictive analytics use churn prediction models that predict customer churn by assessing their propensity of risk to churn. Since these models generate a small prioritized list of potential defectors, they are effective at focusing customer retention marketing programs on the subset of the customer base who are most vulnerable to churn.

The goal of this project is to predict behaviors of churn or not churn using classification methods like Decision Tree, Neural Network and Support Vector machines to help retain customers.

**Data:**

The Data set is taken from kaggle.com

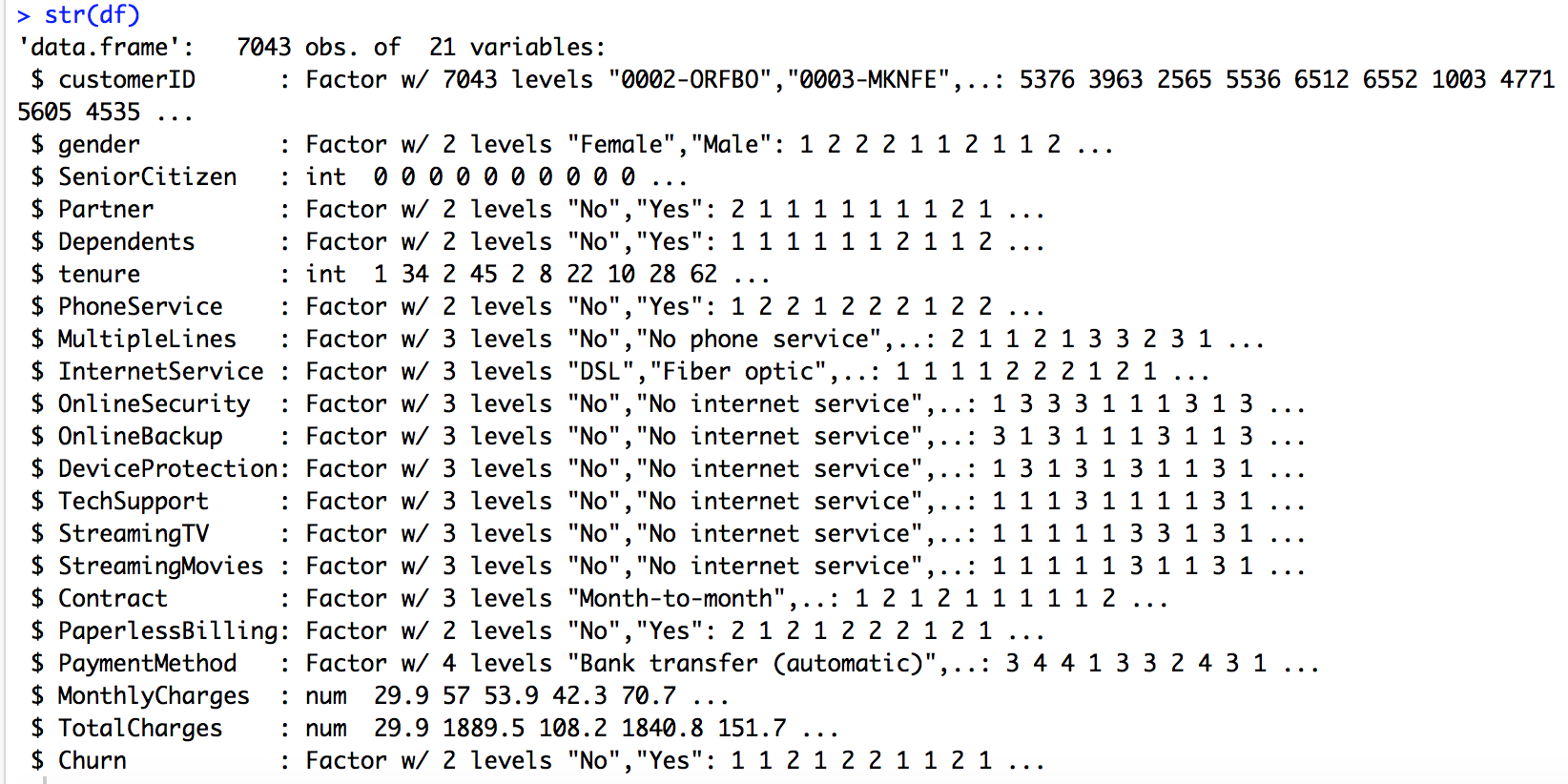
<https://www.kaggle.com/blastchar/telco-customer-churn>

We can analyze all relevant customer data and develop focused customer retention programs by using the dataset.

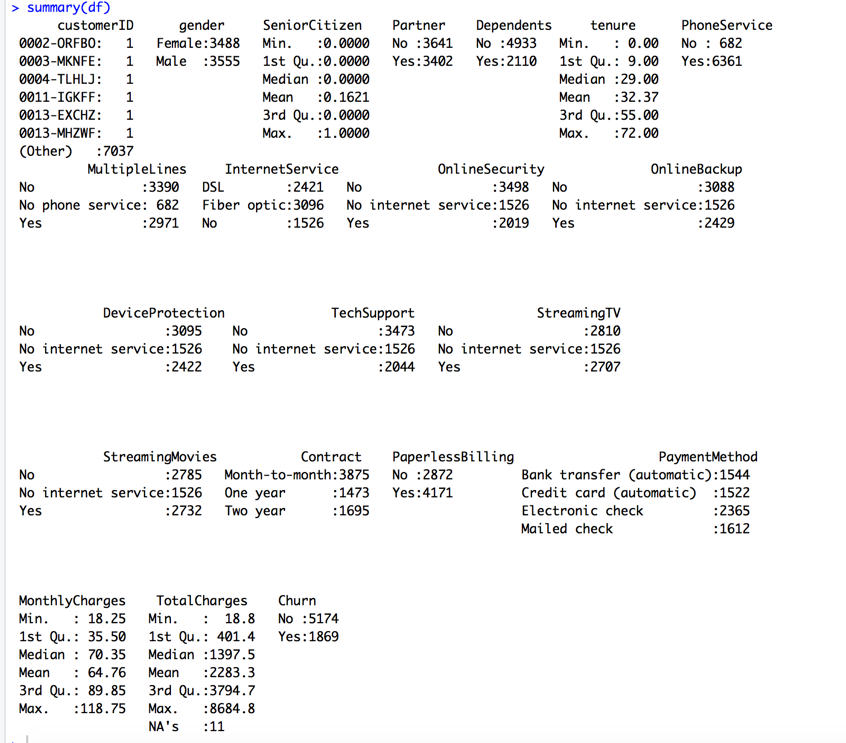
Each row represents a customer; each column contains that customer’s attributes:

* CustomerID
* gender (female, male)
* SeniorCitizen (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)
* PhoneService (Whether the customer has a phone service or not (Yes, No))
* MultipleLines (Whether the customer has multiple lines r not (Yes, No, No phone service)
* InternetService (Customer’s internet service provider (DSL, Fiber optic, No)
* OnlineSecurity (Whether the customer has online security or not (Yes, No, No internet service)
* OnlineBackup (Whether the customer has online backup or not (Yes, No, No internet service)
* DeviceProtection (Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport (Whether the customer has tech support or not (Yes, No, No internet service)
* streamingTV (Whether the customer has streaming TV or not (Yes, No, No internet service)
* streamingMovies (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)
* PaperlessBilling (Whether the customer has paperless billing or not (Yes, No))
* PaymentMethod (The customer’s payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)))
* MonthlyCharges (The amount charged to the customer monthly — numeric)
* TotalCharges (The total amount charged to the customer — numeric)
* Churn ( Whether the customer churned or not (Yes or No))

The raw data contains 7043 rows (customers) and 21 columns (features). The “Churn” column is our target.



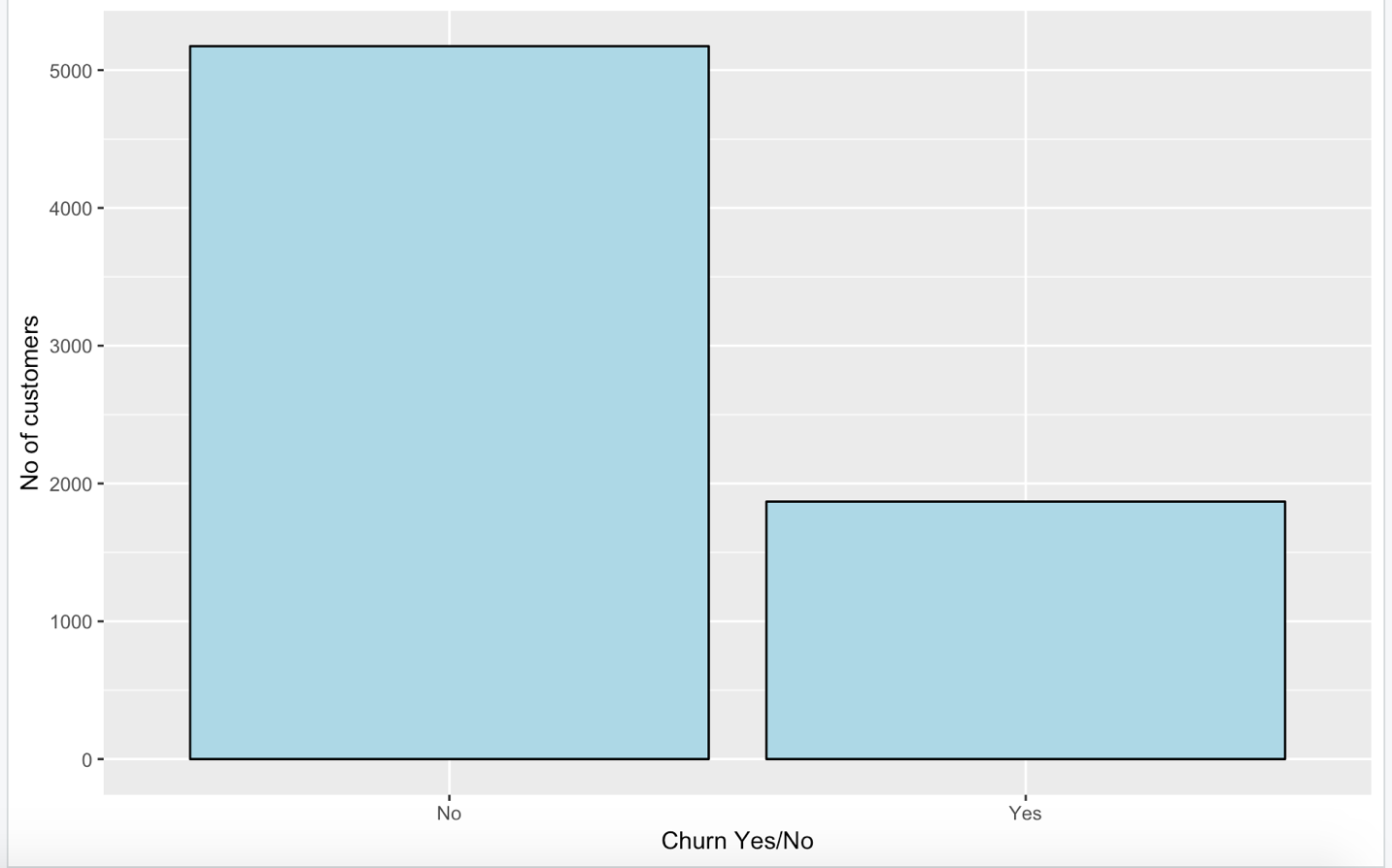
|  |  |  |
| --- | --- | --- |
| Models | Accuracy for Training set | Accuracy for Test set |
| K-Nearest Neighbors Model | Train = 0.904 | Test = 0.907 |
| Random Forest Model | Train=0.907 | Test=0.905 |

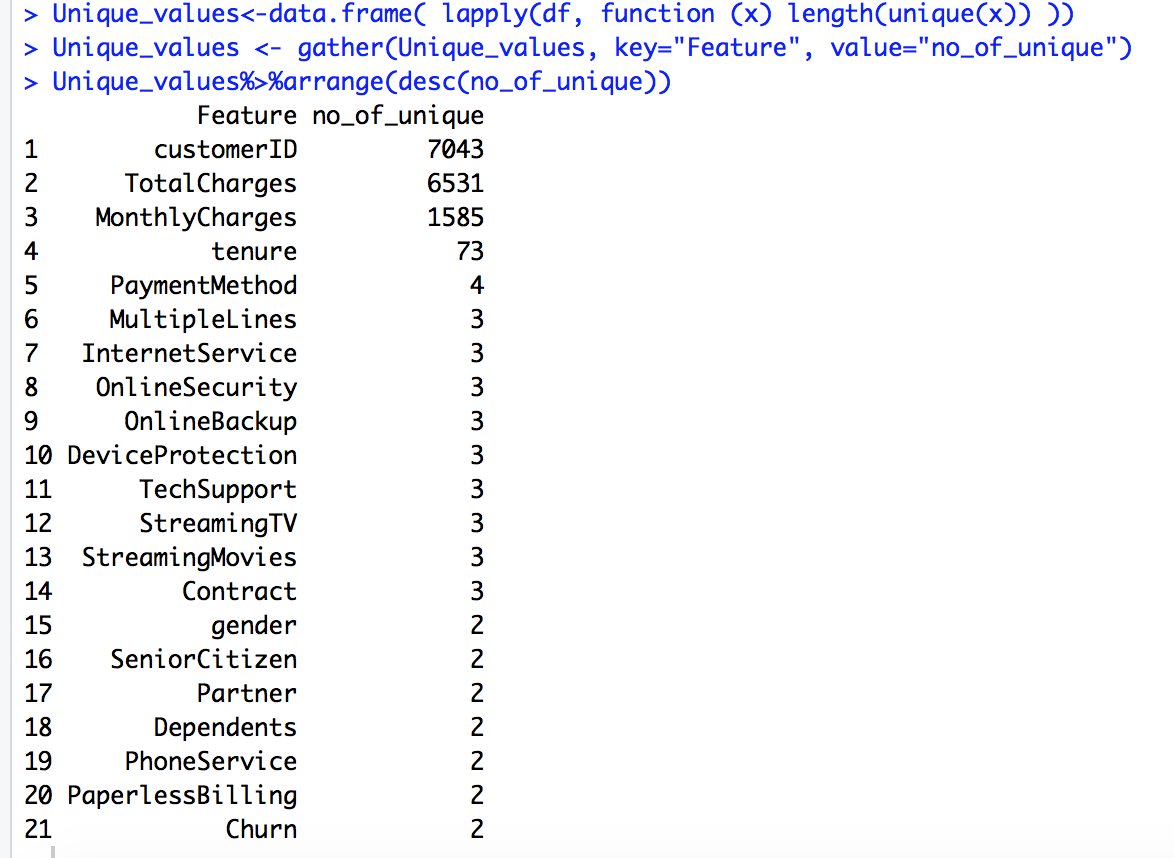


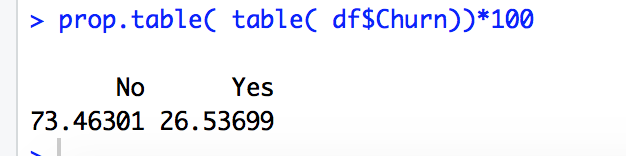
**Data Preprocessing:**

Observations with Missing Values:

Based on the summary, there are 11 missing values in the TotalCharges column, which account for only 0.16% of the total number of observations. So I remove those 11 rows with missing values.







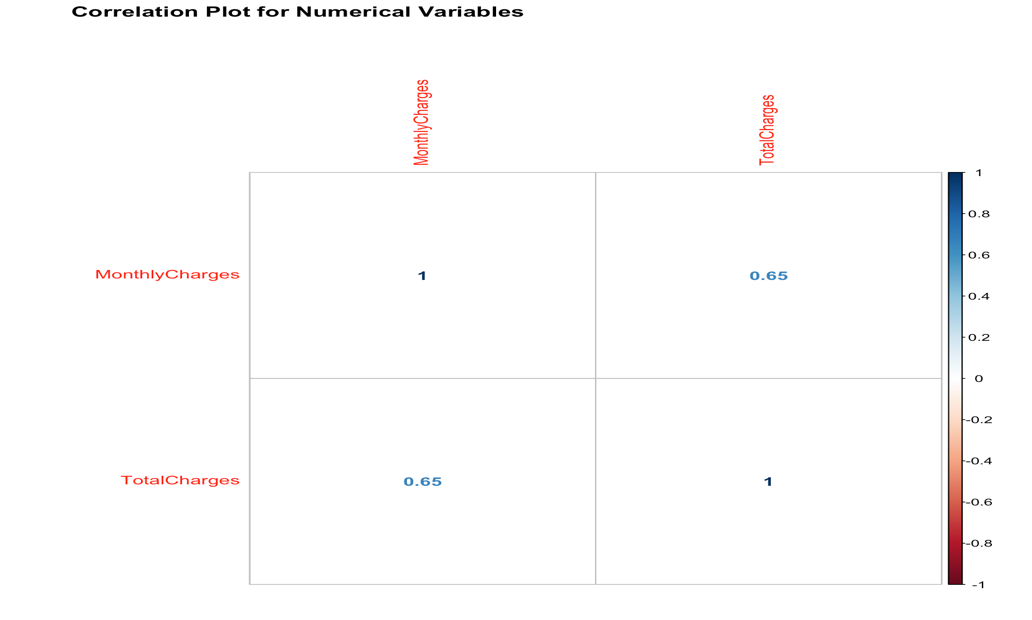
26.5% of customers in this dataset have churned.

**By looking at the variables, there is some wrangling to do.**

1. We will change “No internet service” to “No” for six columns, they are: “OnlineSecurity”, “OnlineBackup”, “DeviceProtection”, “TechSupport”, “streamingTV”, “streamingMovies”.
2. We will change “No phone service” to “No” for column “MultipleLines”.
3. Since the minimum tenure is 1 month and maximum tenure is 72 months, we can group them into five tenure groups: “0–12 Month”, “12–24 Month”, “24–48 Months”, “48–60 Month”, “> 60 Month”
4. Change the values in column “SeniorCitizen” from 0 or 1 to “No” or “Yes”.
5. Remove the columns like customerid and tenure, as we do not need them for the analysis.

### **Exploratory data analysis and feature selection:**

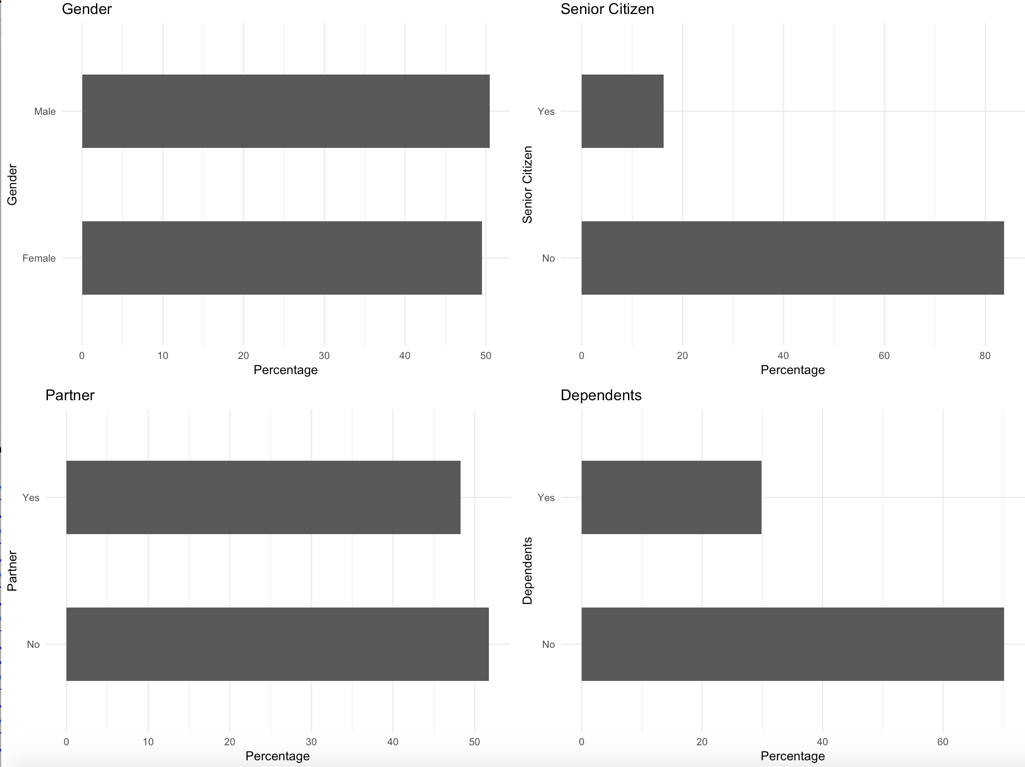
Correlation between numeric variables:



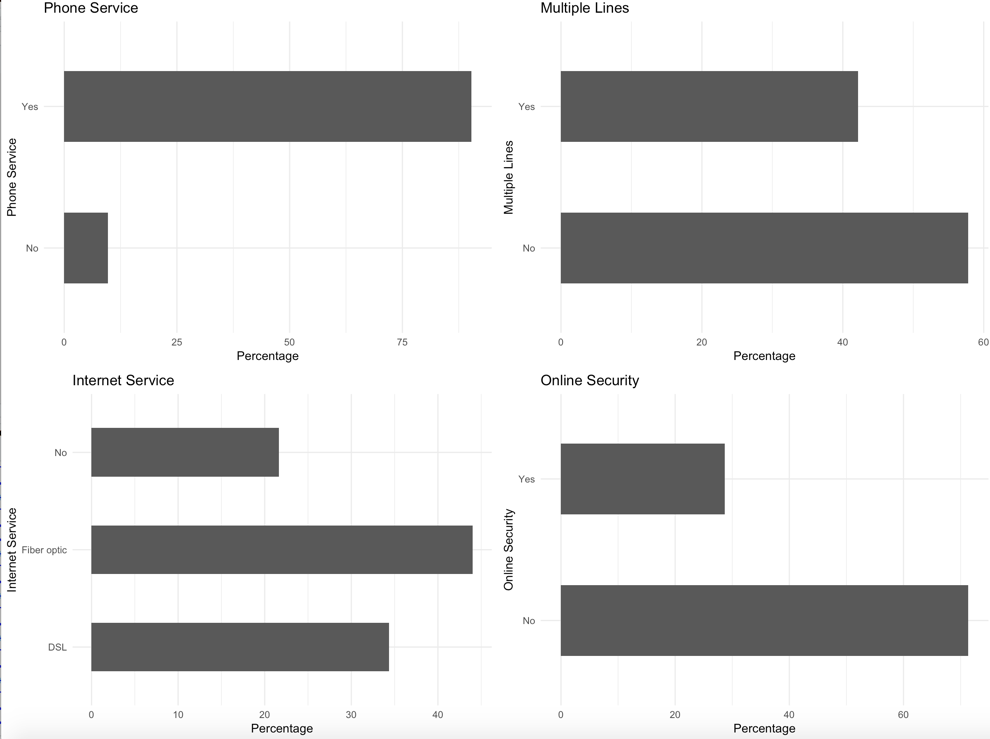
The Monthly Charges and Total Charges are correlated. So one of them can be removed from the model. I will remove Total Charges.

**Bar plots of categorical variables:**

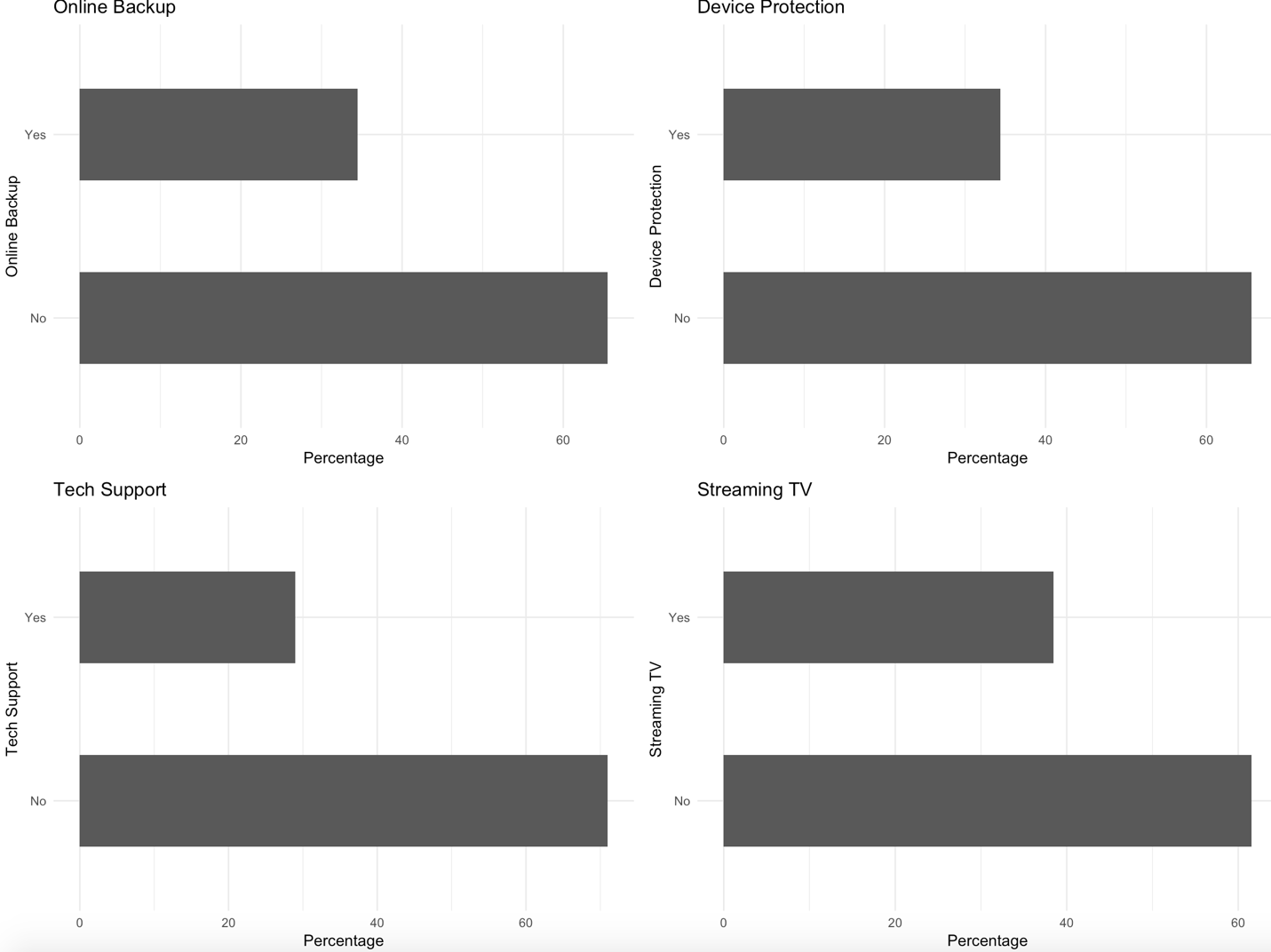
1. Gender, Senior Citizen, Partner and Dependents



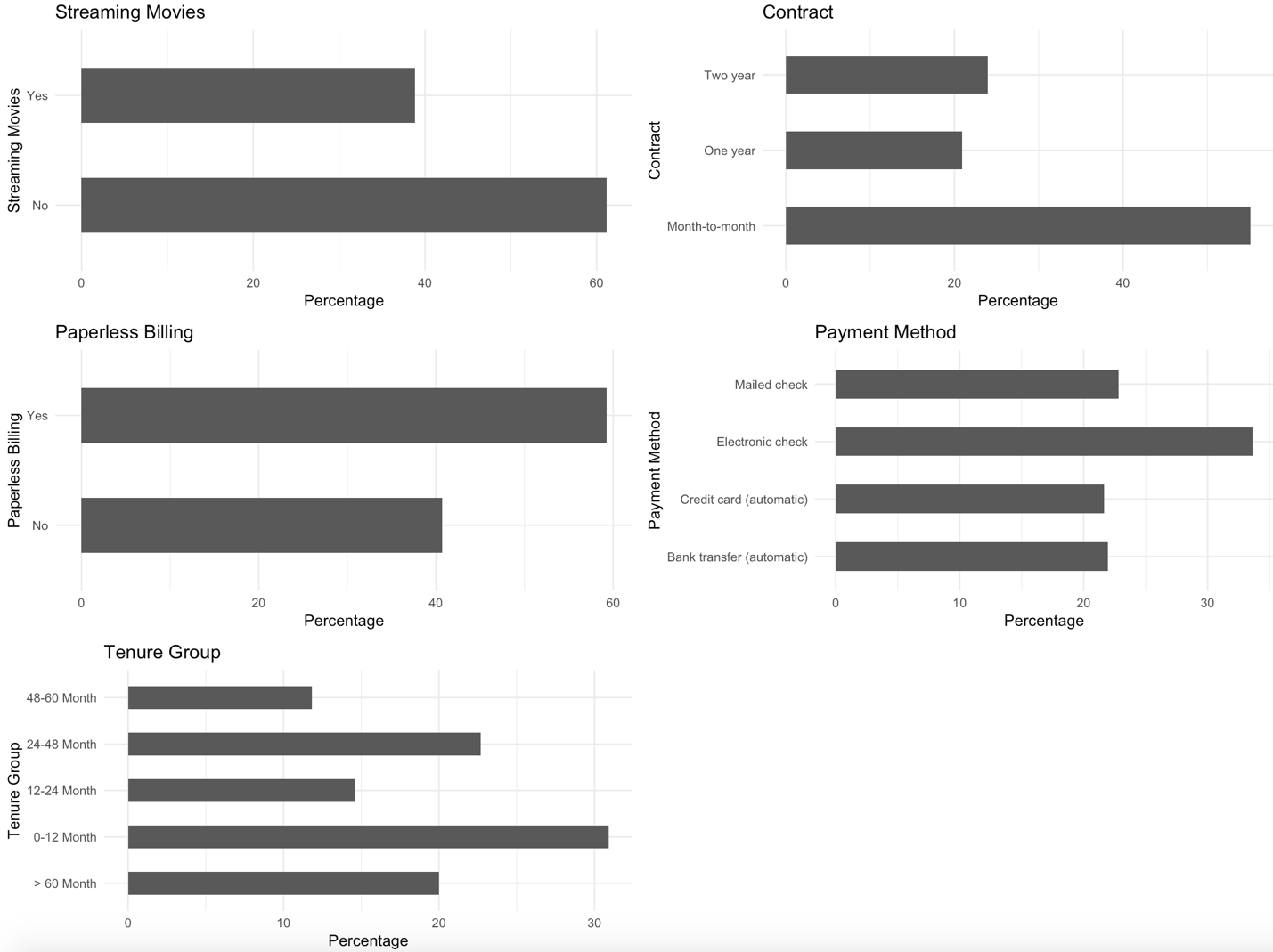
1. Phone Service, Multiple lines, Internet service and Online Security



1. Online Backup, Device Protection, Tech Support and Streaming TV



1. Streaming Movies, Contract, Paperless Billing, Payment Method and Tenure group.



It looks like all of the categorical variables have a reasonably broad distribution, therefore, all of them can be considered for further analysis.

**Classification Methods:**

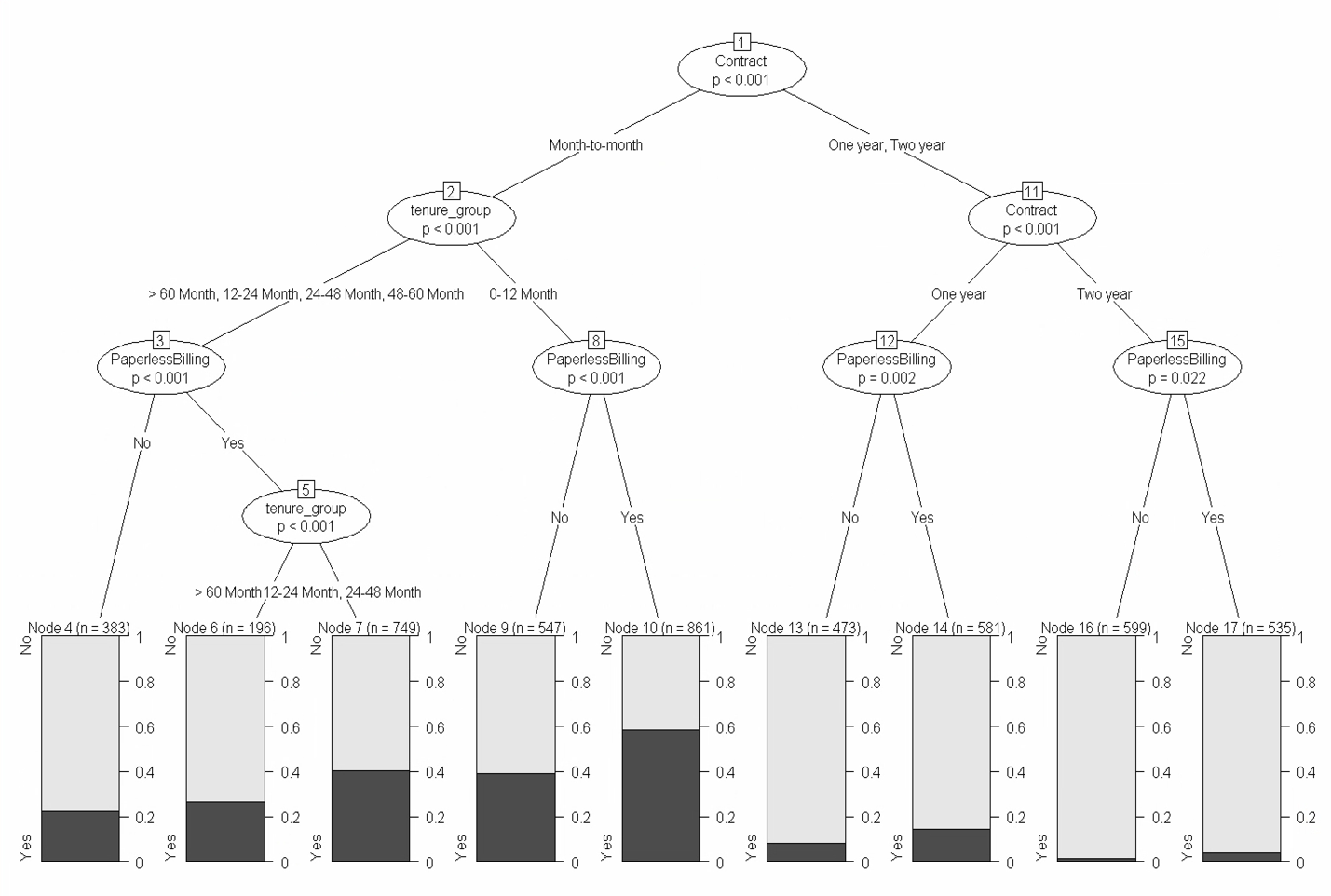
Classification algorithms such as Decision Tree, Neural Networks and Support Vector Machines can be used to predict churn. This algorithms are suitable for this kind of dataset.

Then we can compare their performance to choose the best model

**Decision Tree:**

Decision Tree visualization

For illustration purpose, I will be using only three variables for plotting Decision Trees, they are “Contract”, “tenure\_group” and “PaperlessBilling”.



1. Out of three variables we used, Contract is the most important variable to predict customer churn or not churn.
2. If a customer in a one-year or two-year contract, no matter he (she) has PapelessBilling or not, he (she) is less likely to churn.
3. On the other hand, if a customer is in a month-to-month contract, and in the tenure group of 0–12 month, and using PaperlessBilling, then this customer is more likely to churn.

I have created table with models and their equivalent metrics below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Models\Metrics | Accuracy | Sensitivity | Specificity | Precision |
| Decision Tree | 78.56% | 89.86% | 47.32% | 82.50% |
| Neural Network | 78.51% | 88.95% | 49.64% | 83% |
| SVM | 77.28% | 89.73% | 42.86% | 81.28% |

Precision refers to the percentage of the results which are relevant. The question that this metric answer is of all customers that were labeled as churned, how many actually churned?

Precision and Specificity would be of higher importance in this case.

In order to illustrate recall and precision for each model, lets compute the weighted F-measure. We collect Recall and Precision value from confusion matrix to compute F-measure for each model.

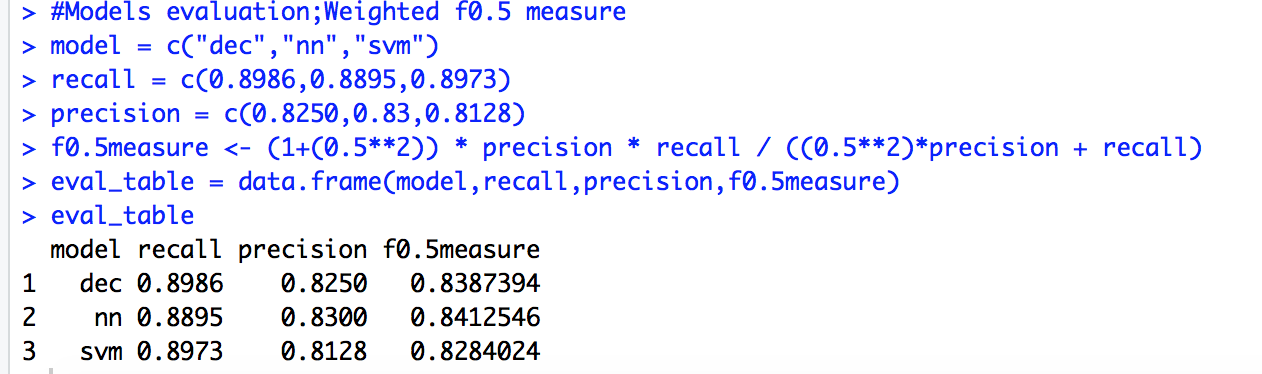
As I am taking Positive class as ‘No’, we should select the model with the highest precision.

**Weighted F0.5 measure**:

(1+β2 )\*precision \* recall /((β2)\*precision + recall)

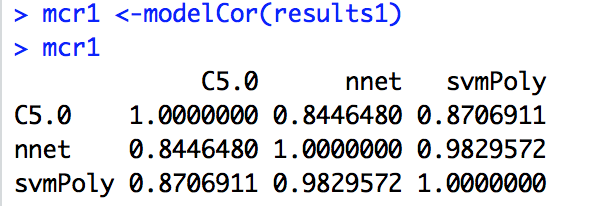
0<β<1 gives more weight to precision while β>1 gives more weight to recall.

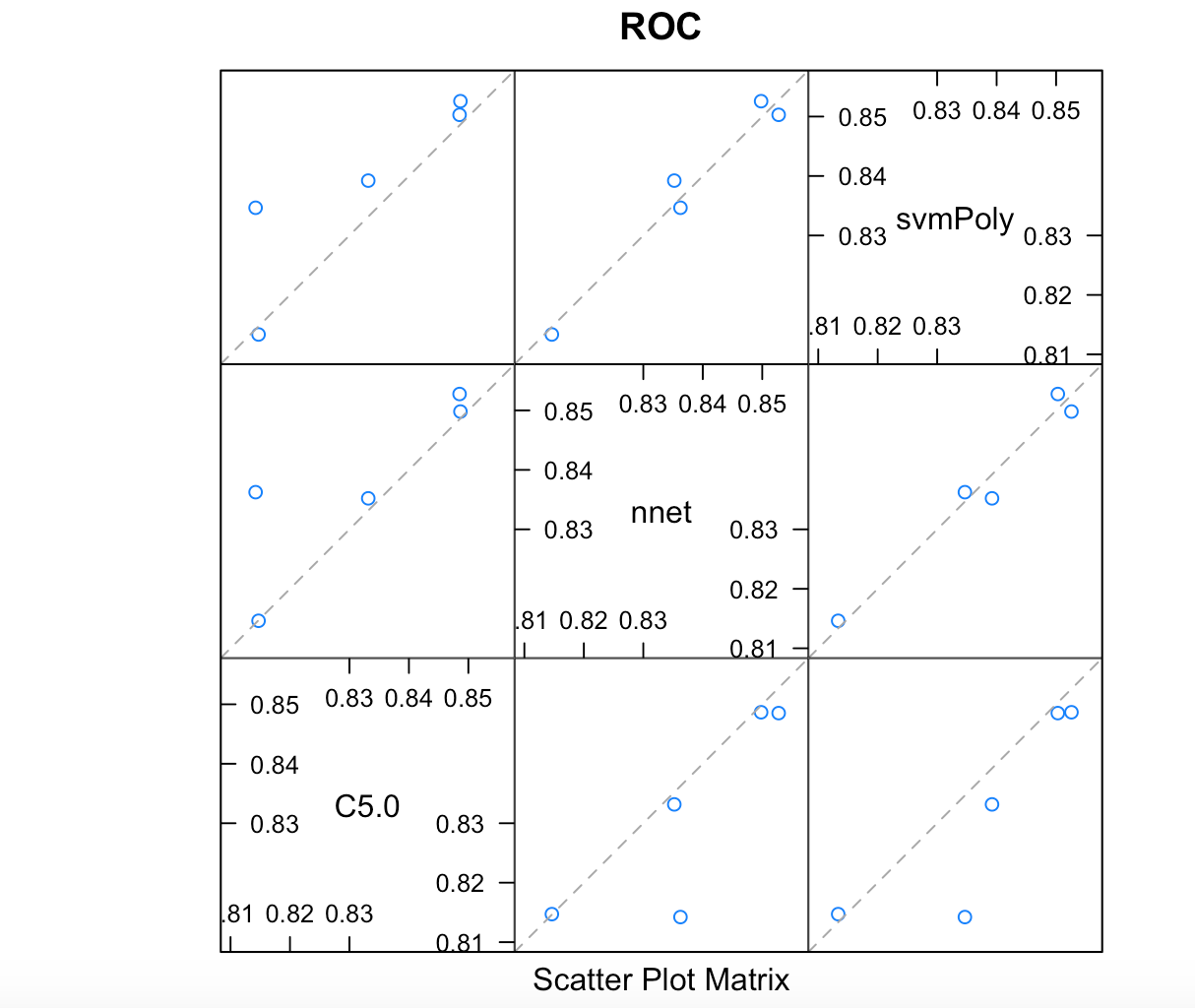
As I want to give more weight to precision, I am assuming β value as 0.5.



Based on the above table, Neural network method is the recommended classification method as it contains highest precision and highest F-measure than other two models.

**Model Correlation:**





From the correlation matrix we can observe that nnet is more positively correlated with svmpoly.

**Summary:**

From the above example, we can see that Decision Tree, Neural Networks and Support vector machine can be used for customer churn analysis for this particular dataset equally fine. The model with highest precision is the Neural Network method.

Below are my few observations:

* Features such as tenure\_group, Contract, PaperlessBilling, MonthlyCharges and InternetService appear to play a role in customer churn.
* Customer who pay with Electronic check are more likely to leave the telco services. It is possible that the customers are not happy with online payement method so company should try to find problems that may exist in this service.
* There does not seem to be a relationship between gender and churn.
* Customers in a month-to-month contract, with PaperlessBilling and are within 12 months tenure, are more likely to churn; On the other hand, customers with one or two year contract, with longer than 12 months tenure, that are not using PaperlessBilling, are less likely to churn.
* Customer, who pay month to month are more likely to leave the telco and so are those customers who pay higher Monthly Charges.

**Analysis and Recommendation:**

The customer needs that should be addressed should be clear. The company must begin to cater to customers that use and expect a higher volume and quality of service. As we can see Contract, MonthlyCharges etc., play important role in customer churn, long term contracts may need to have monthly charges reduced or additional services added at no cost. Month-to-month customers, which make up a large part of the customer base, must be a priority. Since not having internet service is the top contributor, it must be marketed towards customers that do not have it or it should be priced lower. “Power-user” customers need to be oﬀered lower prices, such as packages that include a majority of services at a reduced price.

These suggestions may require additional market and business research, but they are a starting point in reducing customer churn.