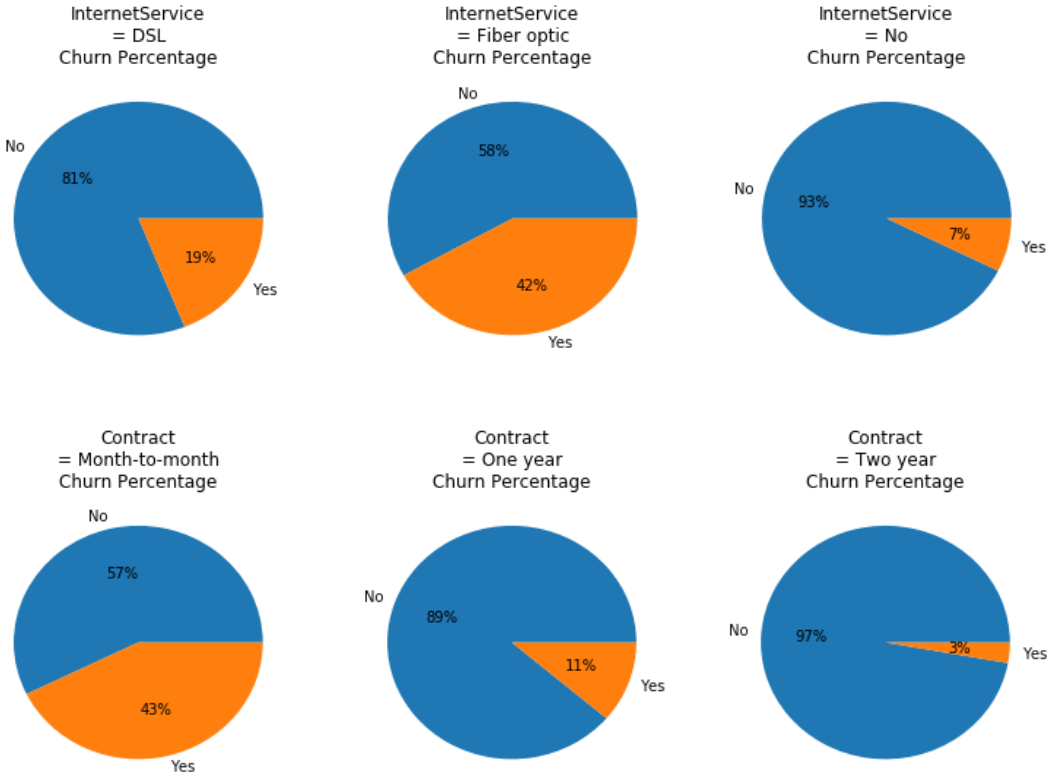
Matric Number: A0164651E HW2 Write-up

1. Descriptive Statistics from EDA *(Refer to Appendix A, page 5)*

From the EDA, we can make a preliminary hypothesis that Contract and InternetService will significantly affect if a customer churns, whereby a customer on Fiber optic will be 6 times more likely to churn than a customer with no InternetService, and a customer on Month-to-month contract is 14 times more like to churn than a customer on Two year contract.



The rest of the plots can be found in *Appendix A, page 4*.

2. Data Pre-Processing and Analysis

The following steps were taken for data cleanup:

a. Managing missing data: TotalCharges field is left blank for customers with tenure = 0. These fields were filled with 0.
b. Correcting data types: TotalCharges was found to be an object type. It was corrected to float type.
c. Checking for and correcting binary variables: eg. 'No internet service' replaced with 'No'

The following steps were taken for data pre-processing

a. Standardization of numerical variables with a standard scaler. $x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j}$
b. Converting binary variables to 0s and 1s so they can be used for calculations
c. Replacing non-binary categorical variables with dummy variables
d. Using a correlation plot, identified and removed 'TotalCharges' and 'MonthlyCharges', as they are highly dependent on other variables *(refer to Appendix B, page 11)*

3. Modelling

Seven models were used. Namely, L1 Logistic Regression, L2 Logistic Regression, Decision Tree (entropy based), Bootstrap Aggregating, Random Forest (entropy based), AdaBoost and Kth Nearest Neighbour. These models were trained and tested on a 70-30 train-test set ratio. They were then tested for accuracy with a 5-folds KFolds method.

3.1 Optimising Accuracy of models

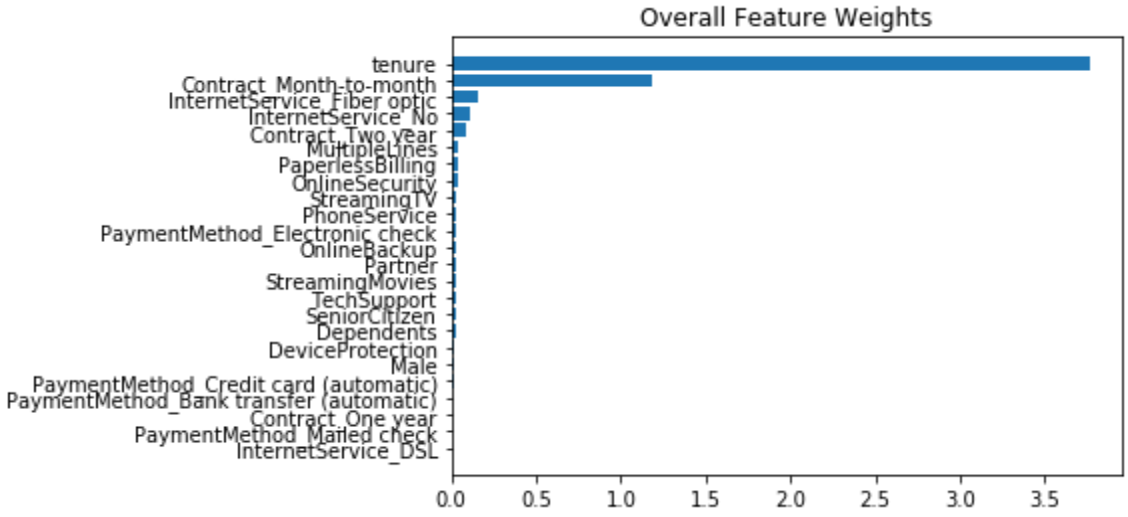
Overfitting arises when a decision tree is excessively dependent on irrelevant features of the training data with the result that its predictive power for unseen instances is reduced. Univariate feature selection was hence done for the Ensemble learning models. The features were first ranked according to their univariate scores, which is based on an F-test estimate of the degree of linear dependency between the feature and churn. A loop was then ran to select the number of top scoring features that gave a model with the highest average cross validation accuracy rate. By pre-pruning, the model may wrongly classify some of the instances in the training set, but the accuracy for the test set may be greater as it avoids overfitting to the training set.

Similarly, a loop was ran to select the number of neighbours, K, for the Kth Nearest Neighbour method that would give a model model with the highest average cross validation accuracy rate, with a preference for an odd K.

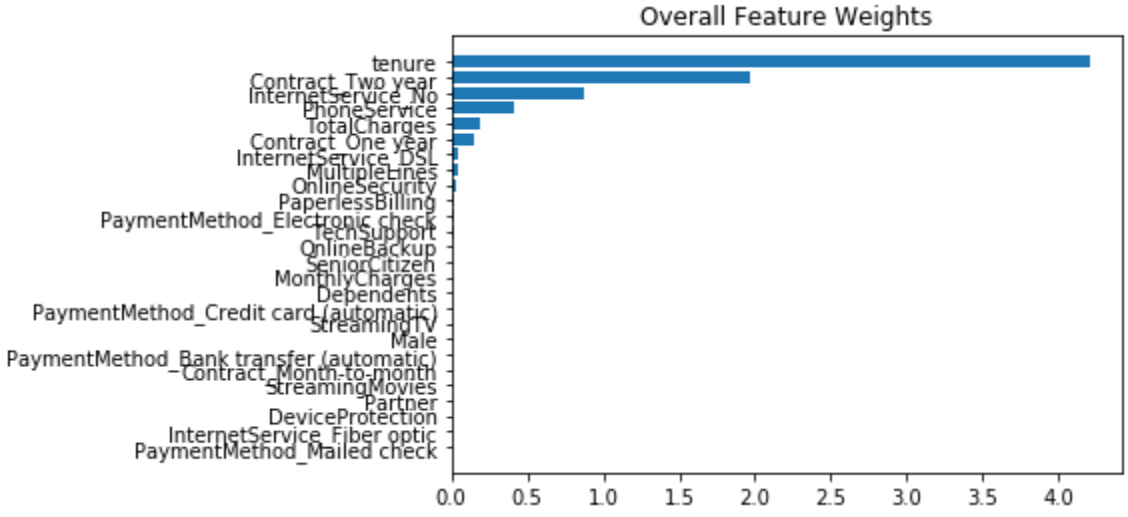
4. Model Results *(Refer to Python Notebook section 5. Model Results)*

The models hit a high average cross validation accuracy of 0.802215 for the Logistic Regression models and a low of 0.774673 for the Decision Tree model.

Finally, the weights assigned to each feature by each model (*Appendix C, page 14)* are then squared, weighted by the model's accuracy and summed. *(Refer to python Notebook section 5.4 for details)*. As the models for both L1 and L2 Logistic Regression are almost identical, we shall omit the weights from L1 Logistic Regression from the weighted sum to avoid double counting. The final feature weights given by the models are as follows:



The feature weights given by the models when all the variables are used (including 'TotalCharges' and 'MonthlyCharges') are as shown: *(Refer to Appendix C, page 12, for detailed results when all the features are used)*



From above, it can be seen that removing highly dependent variables 'TotalCharges' and 'MonthlyCharges' did not cause a change in the top feature, tenure. Furthermore, both runs placed high weightage on InternetService and Contract albeit

different categories. This is in line with our preliminary hypothesis derived from eye-balling the pie charts. These variables are hence the target areas the telco should look into to improve churn rates.

5. Limitations and Other Considerations

1. Target leakage occurs when a model is trained on information that would not be available at the time of prediction and causes a model to over represent its generalization error. Lucky for us, all the features used to train the models can be known at the time of prediction.
2. It is ambiguous whether customers with tenure = 0 should be included in the data for analysis; we do not know if it is possible for customers who are with the telco for less than a month to churn. This is especially important as tenure turns out to be a rather high weighting feature in the final result, and removing these instances may affect the final result.
3. A couple of variables such as Gender and PaymentMethod have a weighting very close to zero. We can probably remove these variables with a recursive feature elimination and check if there are any changes in the weightings.
4. Number of neighbours, K, for the KNN Classification is on tested up till K=80. It might be possible for a K>80 to give a higher cross validation accuracy. However, it is too computationally intensive to calculate all 4930 possible accuracy values.

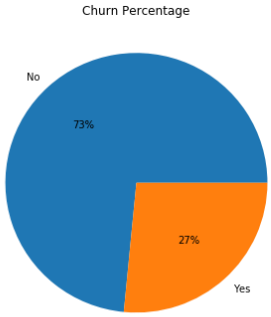
6. Room for improvement/ Other data that could have helped

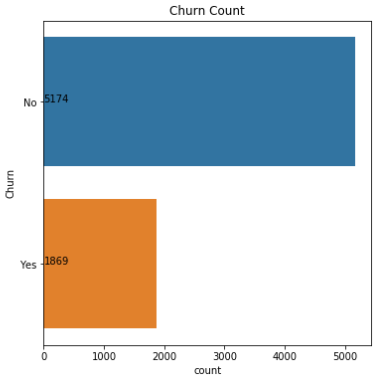
Other factors such as the income level of the customer may have an effect on whether a customer churns. However, customer income data may be difficult for a telco to obtain.

Factors such as when is a customer's contract end may have an impact on whether a customer churns in a particular month. The telco should have this data ready and can incorporate it for better predictions.

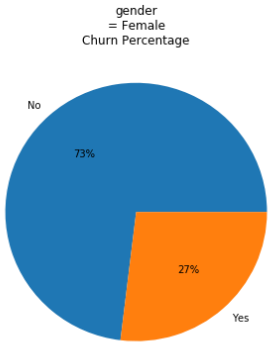
Appendix

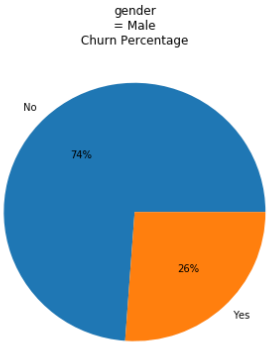
Appendix A

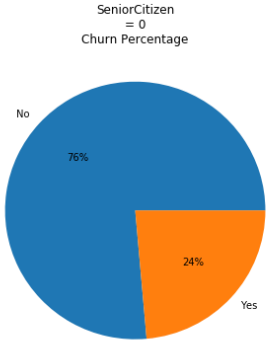


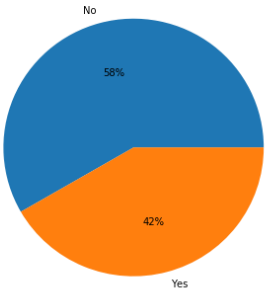


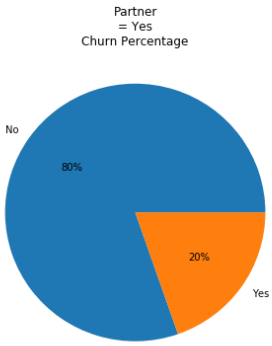
Pie Charts of each Categorical Variable against Churn

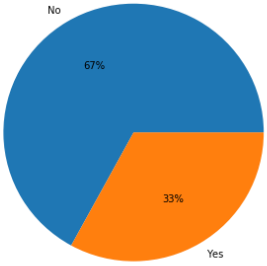




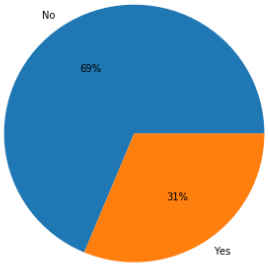




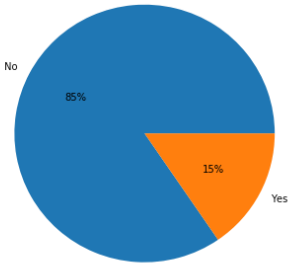




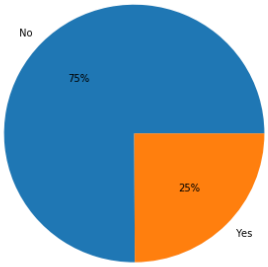
Dependents
= No
Churn Percentage



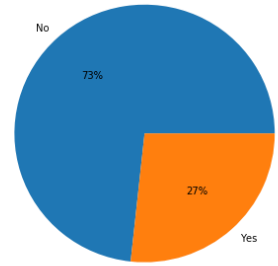
Dependents
= Yes
Churn Percentage



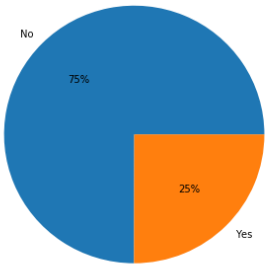
PhoneService
= No
Churn Percentage



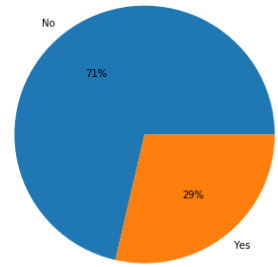
PhoneService
= Yes
Churn Percentage



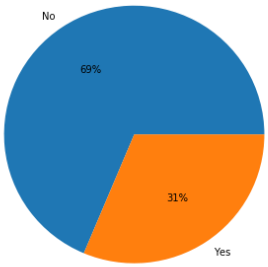
MultipleLines
= No
Churn Percentage



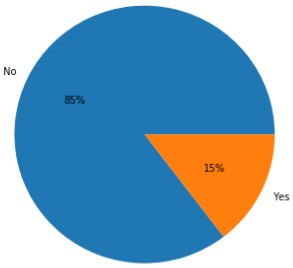
MultipleLines
= Yes
Churn Percentage



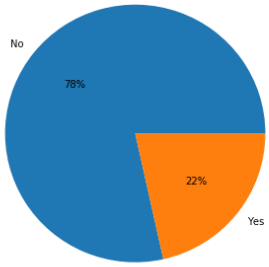
OnlineSecurity
= No
Churn Percentage



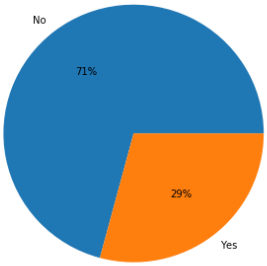
OnlineSecurity
= Yes
Churn Percentage



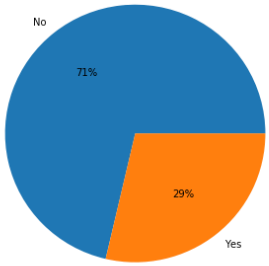
OnlineBackup
= Yes
Churn Percentage



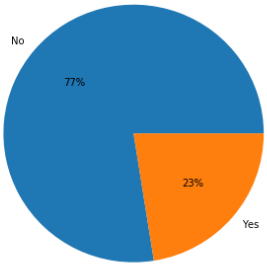
OnlineBackup
= No
Churn Percentage



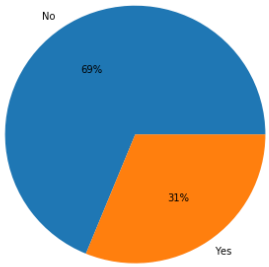
DeviceProtection
= No
Churn Percentage



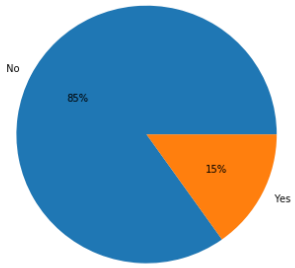
DeviceProtection
= Yes
Churn Percentage



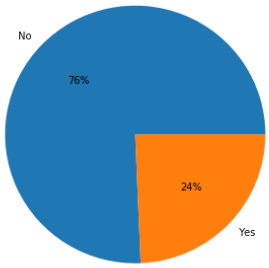
TechSupport
= No
Churn Percentage



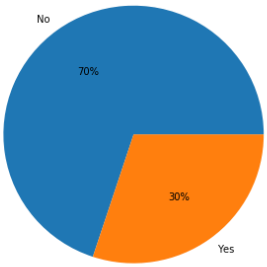
TechSupport
= Yes
Churn Percentage

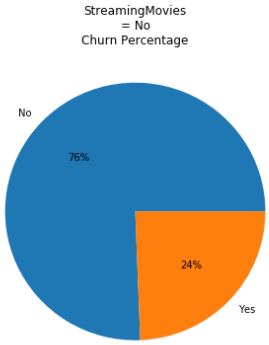


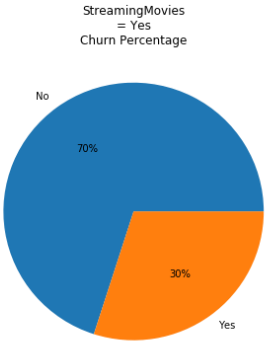
StreamingTV
= No
Churn Percentage

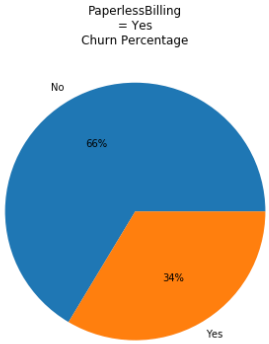


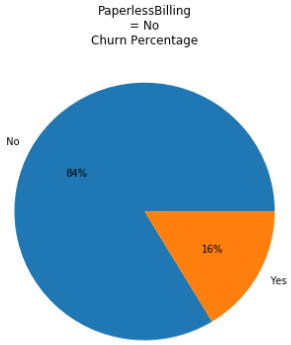
StreamingTV
= Yes
Churn Percentage

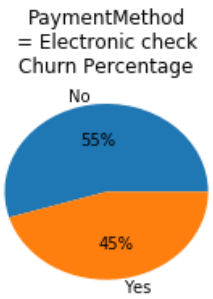


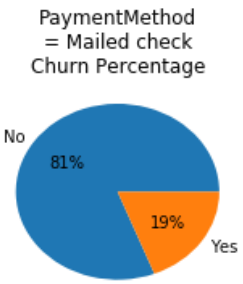

StreamingMovies
= No
Churn Percentage
No
76%
24%
Yes

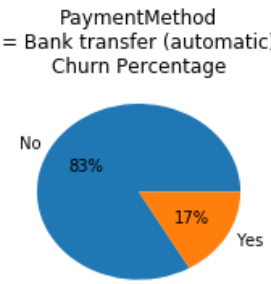

StreamingMovies
= Yes
Churn Percentage
No
70%
30%
Yes

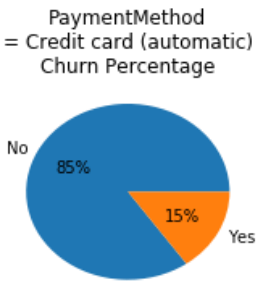

PaperlessBilling
= Yes
Churn Percentage
No
66%
34%
Yes


PaperlessBilling
= No
Churn Percentage
No
84%
16%
Yes

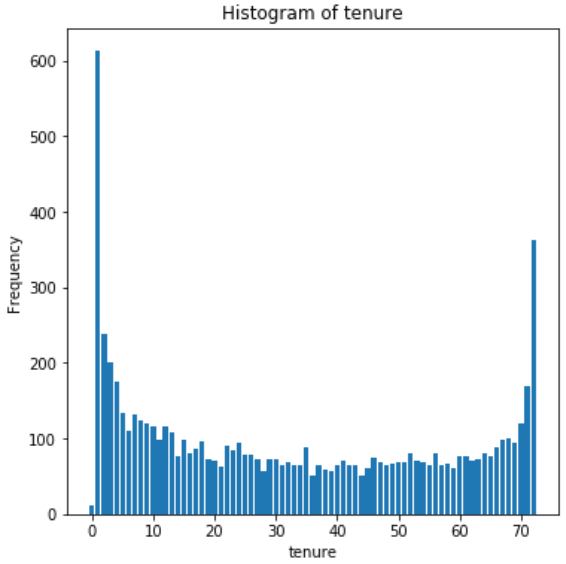

PaymentMethod
= Electronic check
Churn Percentage
No
55%
45%
Yes

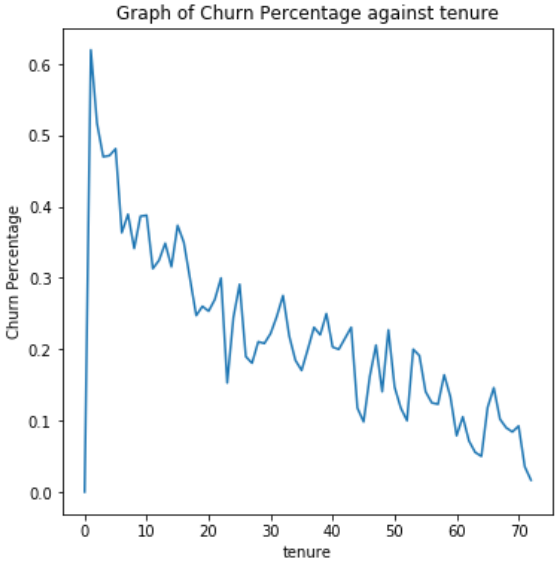

PaymentMethod
= Mailed check
Churn Percentage
No
81%
19%
Yes

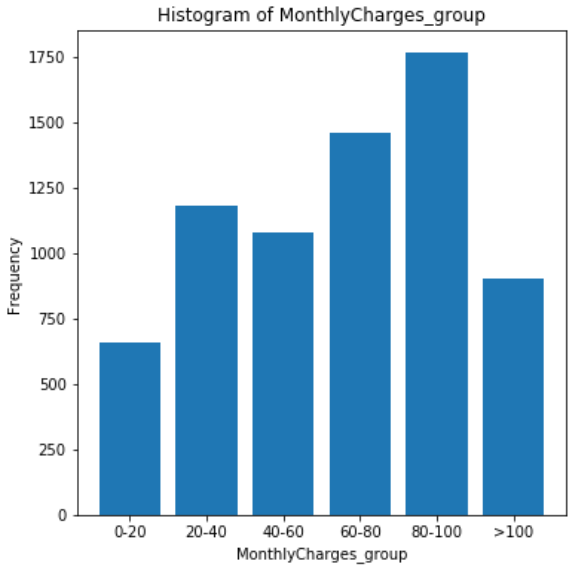

PaymentMethod
= Bank transfer (automatic)
Churn Percentage
No
83%
17%
Yes

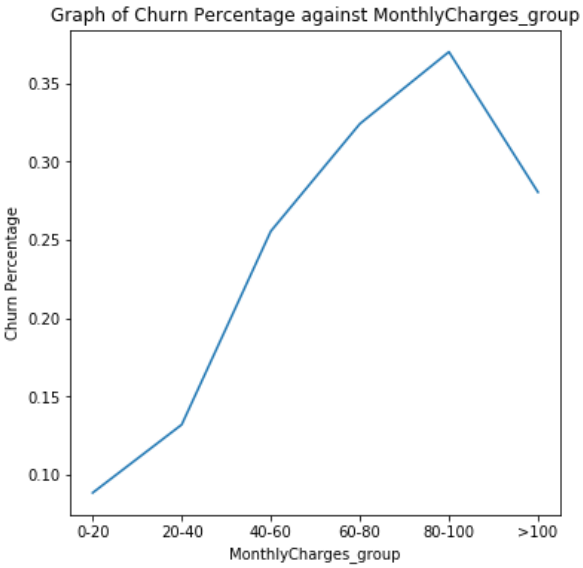

PaymentMethod
= Credit card (automatic)
Churn Percentage
No
85%
15%
Yes

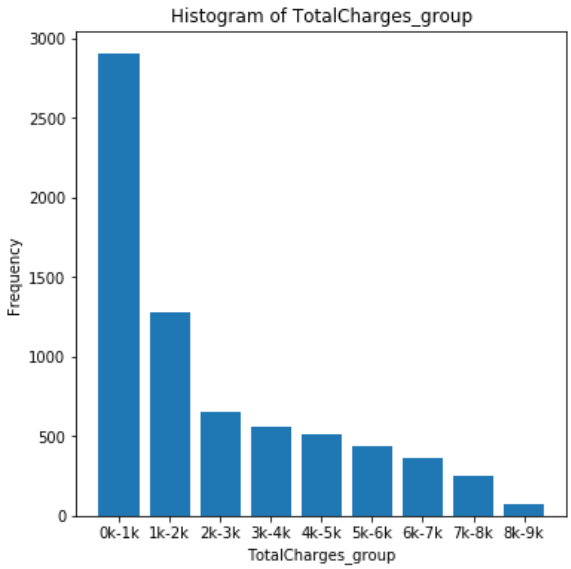
Charts for Numerical Variables

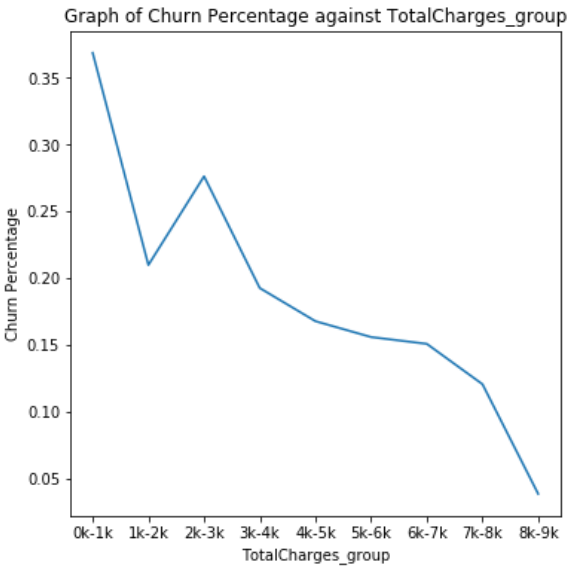






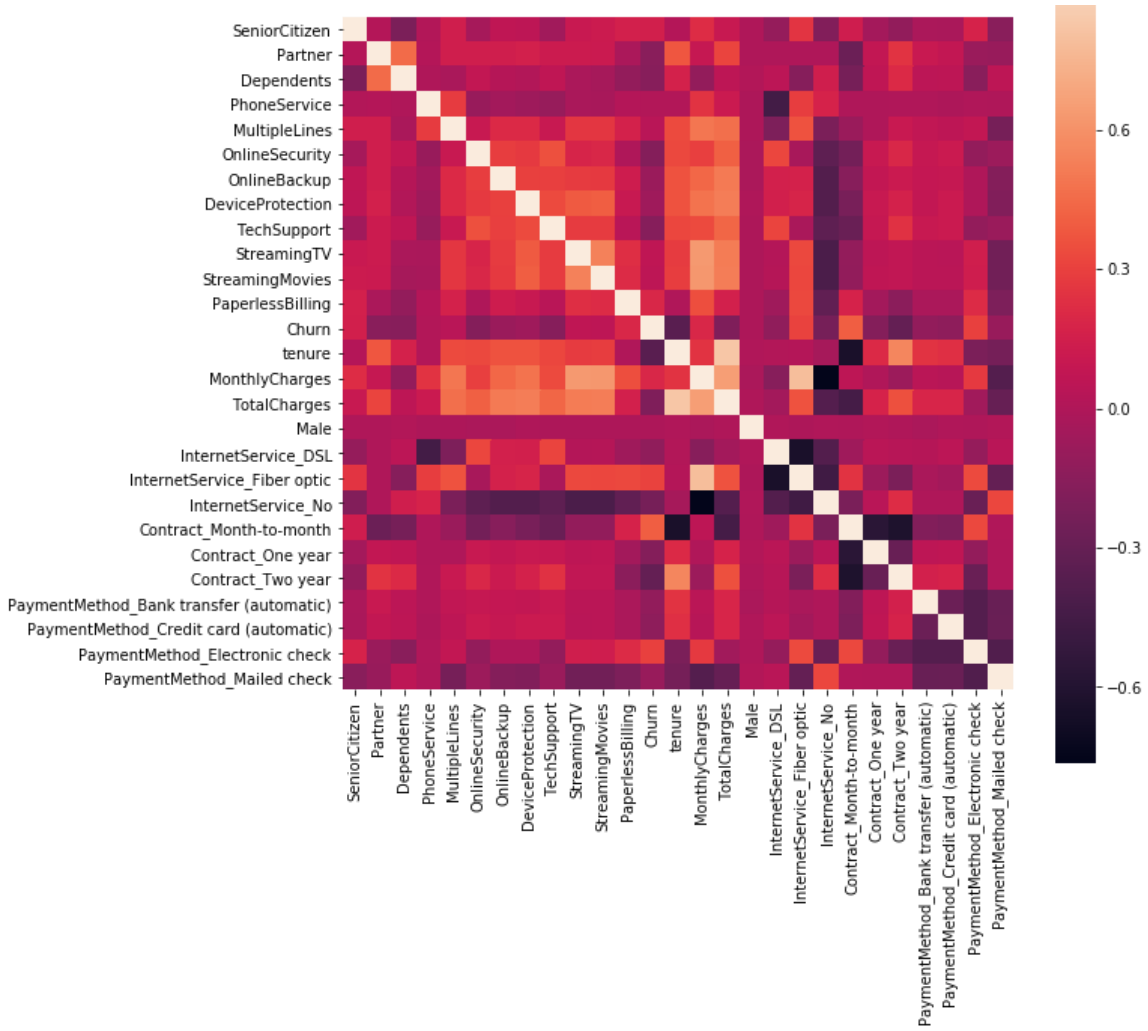






	SeniorCitizen	tenure	MonthlyCharges	TotalCharges
count	7043.000000	7043.000000	7043.000000	7043.000000
mean	0.162147	32.371149	64.761692	2279.734304
std	0.368612	24.559481	30.090047	2266.794470
min	0.000000	0.000000	18.250000	0.000000
25%	0.000000	9.000000	35.500000	398.550000
50%	0.000000	29.000000	70.350000	1394.550000
75%	0.000000	55.000000	89.850000	3786.600000
max	1.000000	72.000000	118.750000	8684.800000

Appendix B: Correlation between variables



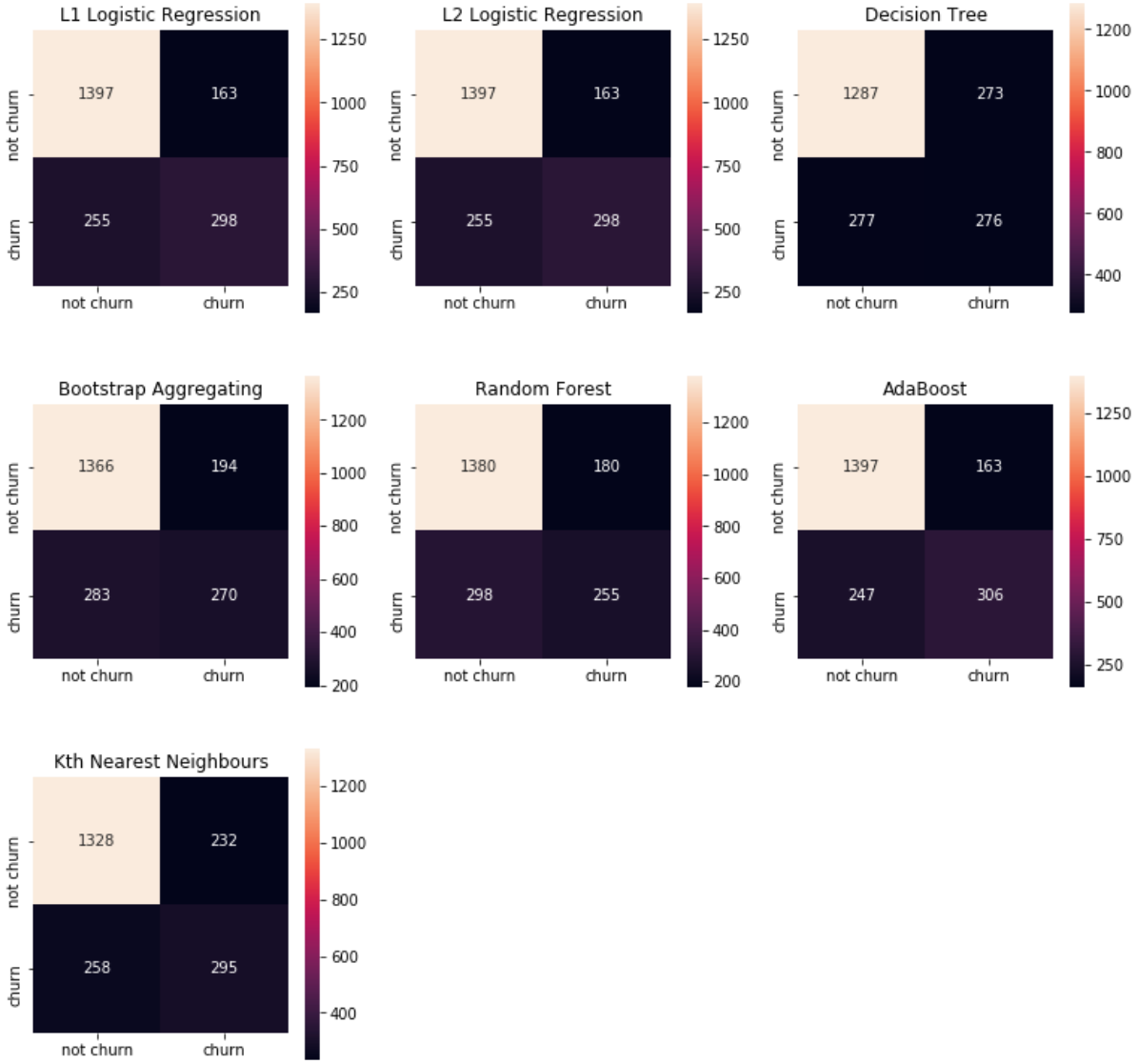

```
0.8261783979502383 = cor(tenure , TotalCharges)
0.787065528472674 = cor(MonthlyCharges , InternetService_Fiber optic)
```

We shall remove (either tenure or TotalCharges) and (either MonthlyCharges or InternetService_Fibre optic). The variable to remove will be the variable with the greatest number of other variables they are highly correlated to, as follows:

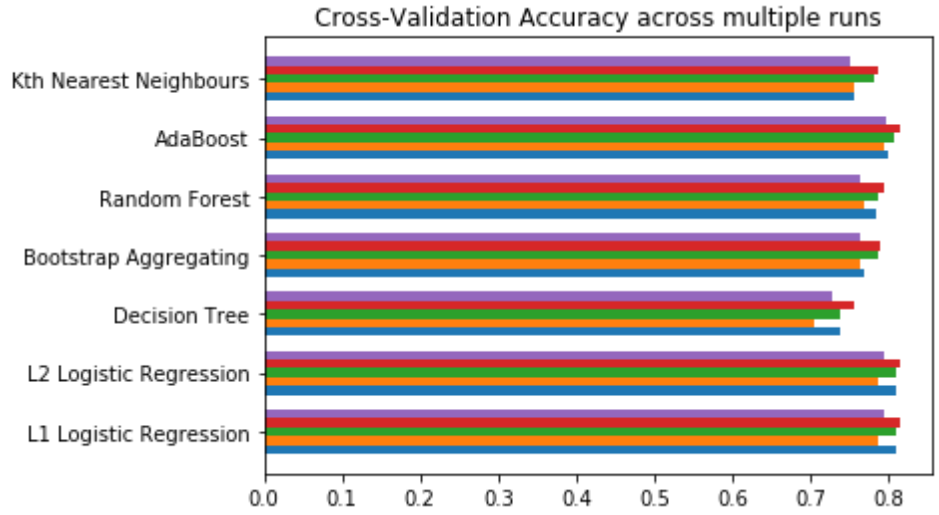
```
0.8261783979502383 = cor(tenure , TotalCharges)
tenure has a correlation coefficient >0.6 with 1 other variables

0.8261783979502383 = cor(TotalCharges , tenure)
0.6511738315787867 = cor(TotalCharges , MonthlyCharges)
TotalCharges has a correlation coefficient >0.6 with 2 other variables

0.6296031609781186 = cor(MonthlyCharges , StreamingTV)
0.6274288843898873 = cor(MonthlyCharges , StreamingMovies)
0.6511738315787867 = cor(MonthlyCharges , TotalCharges)
0.787065528472674 = cor(MonthlyCharges , InternetService_Fiber optic)
MonthlyCharges has a correlation coefficient >0.6 with 4 other variables

0.787065528472674 = cor(InternetService_Fiber optic , MonthlyCharges)
InternetService_Fiber optic has a correlation coefficient >0.6 with 1 other var
iables
```

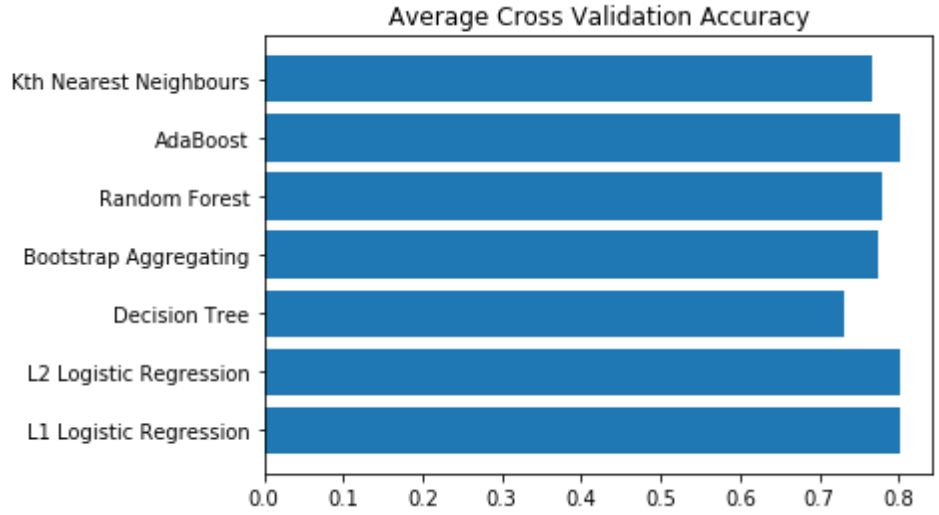
Appendix C: Model Results when all the features are used

Confusion Matrices when all the features are used



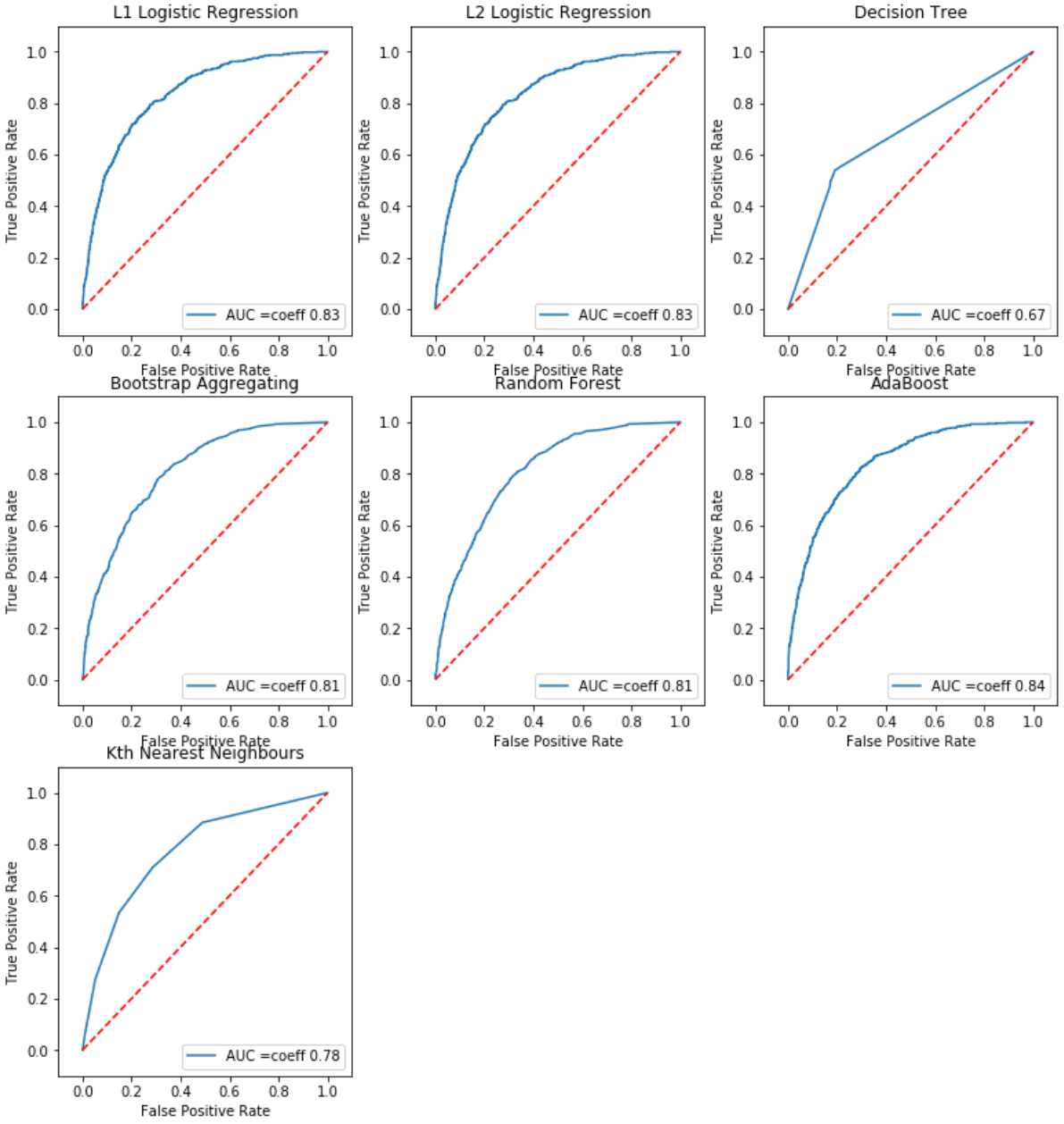
Model Accuracies when all the features are used



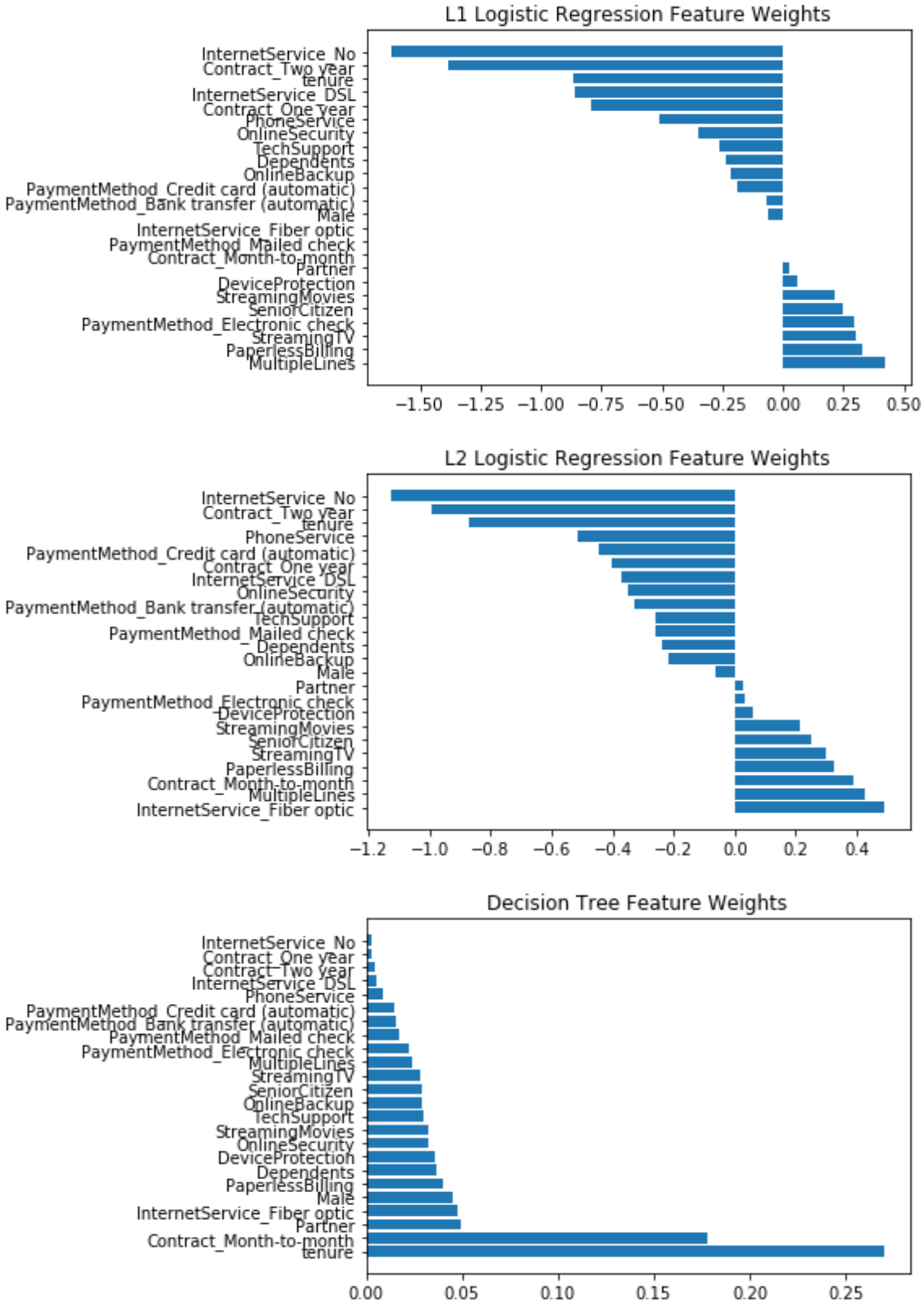


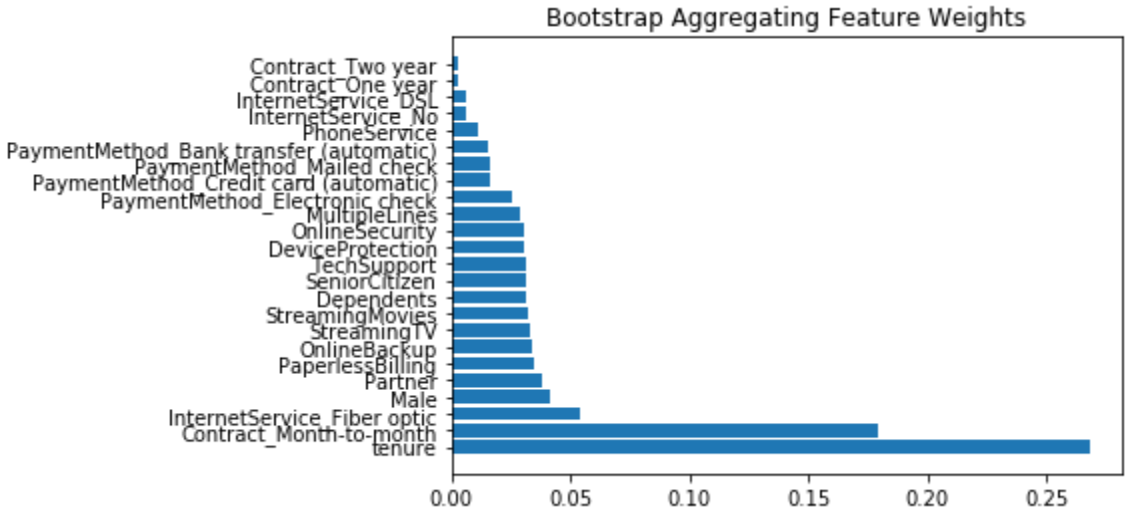
Model	Average Cross Validation Accuracy
L1 Logistic Regression	0.802215
L2 Logistic Regression	0.802215
Decision Tree	0.732219
Bootstrap Aggregating	0.774102
Random Forest	0.779355
AdaBoost	0.802784
Kth Nearest Neighbours	0.765868

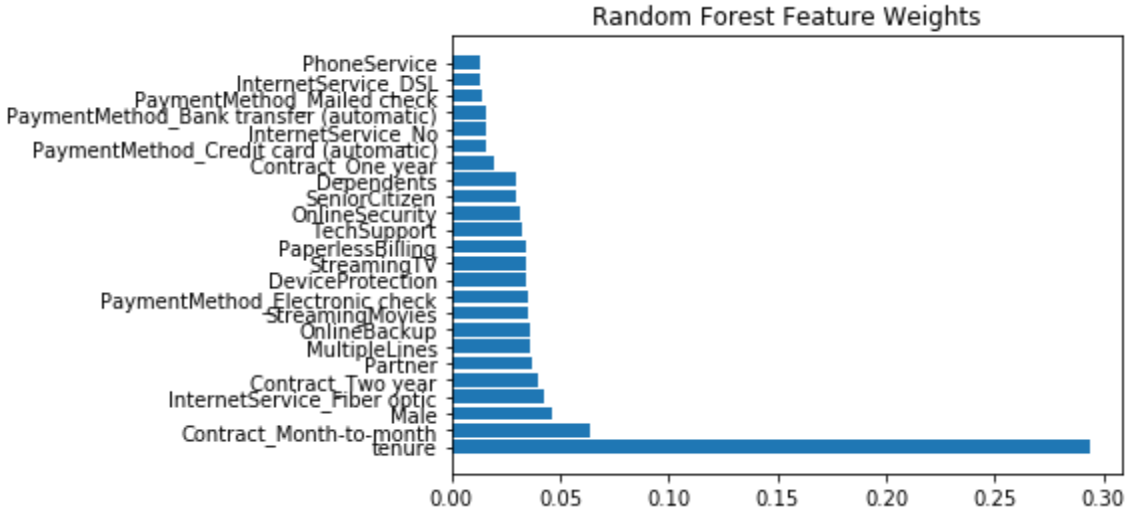
ROC Curves when all the features are used

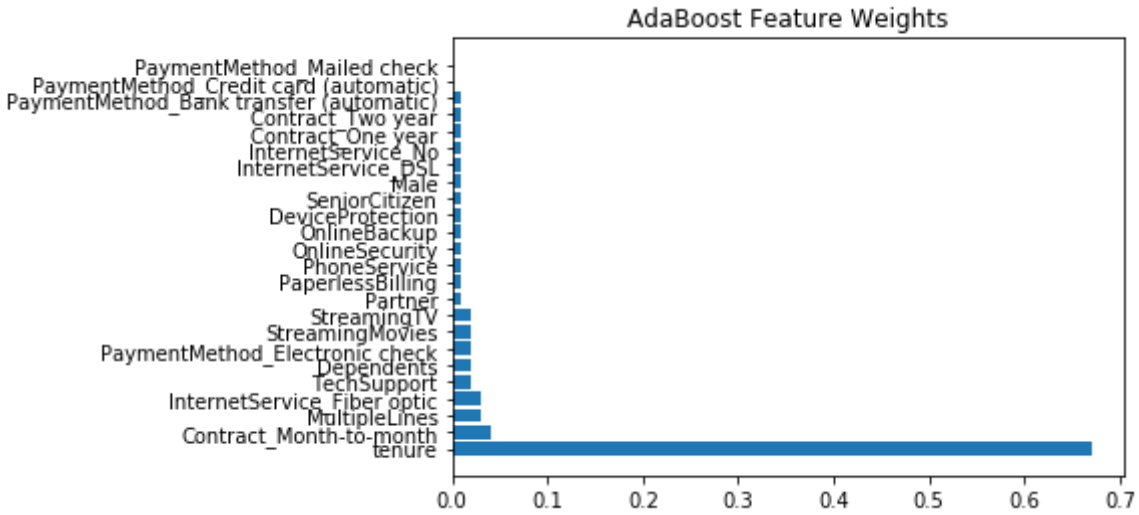


Feature Weights when all the features are used




Bootstrap Aggregating Feature Weights
Contract_Two year
Contract_One year
InternetService_DSL
InternetService_No
PhoneService
PaymentMethod_Bank transfer (automatic)
PaymentMethod_Mailed check
PaymentMethod_Credit card (automatic)
PaymentMethod_Electronic check
MultipleLines
OnlineSecurity
DeviceProtection
TechSupport
SeniorCitizen
Dependents
StreamingMovies
StreamingTV
OnlineBackup
PaperlessBilling
Partner
Male
InternetService_Fiber optic
Contract_Month-to-month
tenure
0.00
0.05
0.10
0.15
0.20
0.25


Random Forest Feature Weights
PhoneService
InternetService_DSL
PaymentMethod_Mailed check
PaymentMethod_Bank transfer (automatic)
InternetService_No
PaymentMethod_Credit card (automatic)
Contract_One year
Dependents
SeniorCitizen
OnlineSecurity
TechSupport
PaperlessBilling
StreamingTV
DeviceProtection
PaymentMethod_Electronic check
StreamingMovies
OnlineBackup
MultipleLines
Partner
Contract_Two year
InternetService_Fiber optic
Male
Contract_Month-to-month
tenure
0.00
0.05
0.10
0.15
0.20
0.25
0.30


AdaBoost Feature Weights
PaymentMethod_Mailed check
PaymentMethod_Credit card (automatic)
PaymentMethod_Bank transfer (automatic)
Contract_Two year
Contract_One year
InternetService_No
InternetService_DSL
Male
SeniorCitizen
DeviceProtection
OnlineBackup
OnlineSecurity
PhoneService
PaperlessBilling
Partner
StreamingTV
StreamingMovies
PaymentMethod_Electronic check
Dependents
TechSupport
InternetService_Fiber optic
MultipleLines
Contract_Month-to-month
tenure
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7