Customer Churn – Telecommunication Industry

Model Selection and Business Implications

Exploratory analysis provided an initial view of the variables (such as fiber optic service and the seniority of citizens) that are associated with high churn rates. The variable importance according to our first model – logistic regression – highlighted not only the variables that are positively related but also those that have a weak (gender and partner) or a negative relation (longer tenures, longer contracts, and tech support) with churn.

For the model selection, it is important to look beyond the overall accuracy of the models. The selection process is driven by the worst case scenario, which is not being able to identify the customer who can churn. Hence, sensitivity (ability to identify true positives) is more important than the overall accuracy of the model.

This approach becomes even more important in the datasets with very low instances of churn: in such cases, the model can achieve accuracy by accurately predicting the customers who would not churn but can have poor performance in identifying true positives.

In our case, logistic regression is superior than the other models in terms of both accuracy and sensitivity; thus, we will select logistic regression on similar datasets to predict whether a customer is about to churn away or not.

In the absence of logistic regression, we would have preferred decision trees to KNN, though KNN has better accuracy – because the decision tree has better sensitivity.

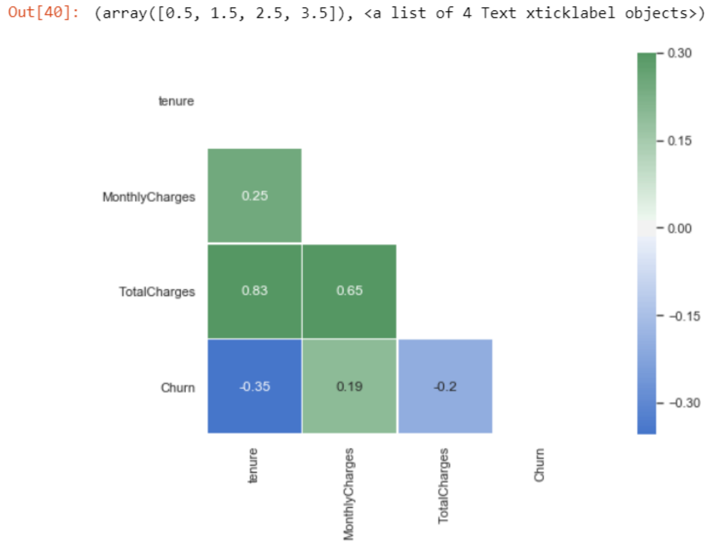
It is important to note that while we are pursuing better sensitivity, we can't totally ignore specificity (ability to identify true negatives). Poor specificity can lead to expenditure on attempts to retain the customers we would have retained them even otherwise (this is the impact of false positives). At what point does the marketing/product teams start focusing on sensitivity is a managerial decision, which will primarily be driven by budget allocation.

Finally, these models can only uncover the areas of concern; it's the managerial interpretations that drive the actions. For instance, strong positive association of fiber optic services with high churn possibility and high price of fiber optic services can't be used to draw the conclusion that high price of the fiber optic services is the cause of the problem. Other possibilities such as poor performance of the product itself or the presence of the product in the wrong market segments can also be the reasons behind customer churn.

Data Source: https://www.kaggle.com/blastchar/telco-customer-churn

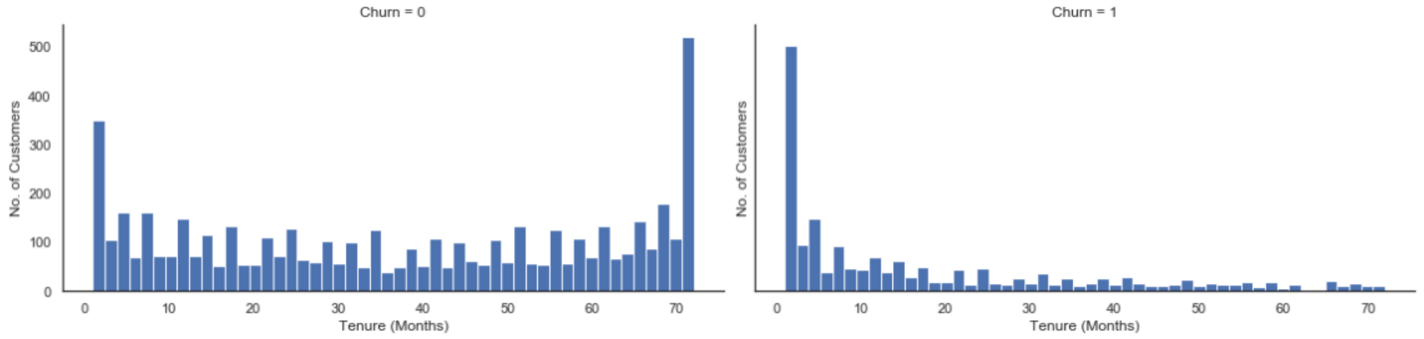
Exploratory Data Analysis

1. Correlation matrix enabled us to know the correlation between different variables. Here, as we observed that 'Total Charges' has a strong positive correlation with other independent variables 'Monthly Charges' and 'Tenure', we removed Total Charges from our model.

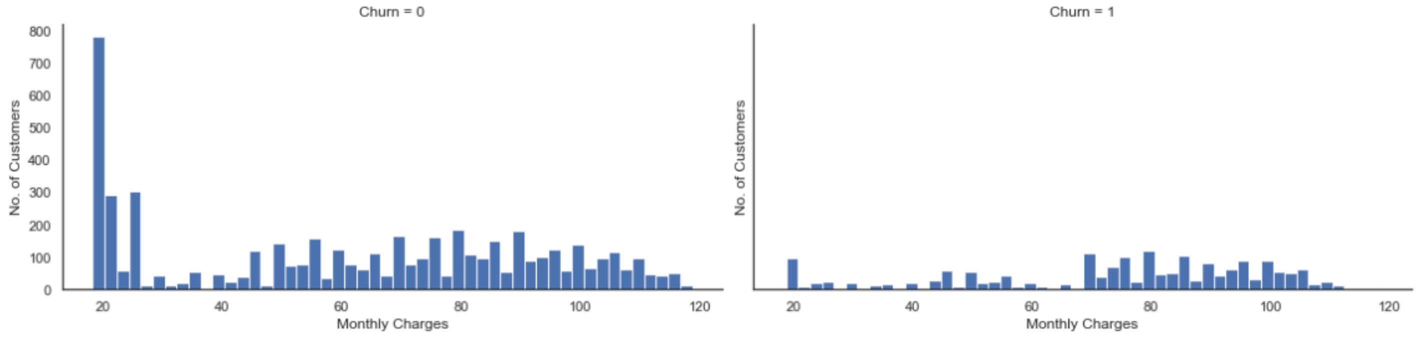


```
Out[40]: (array([0.5, 1.5, 2.5, 3.5]), <a list of 4 Text xticklabel objects>)
```

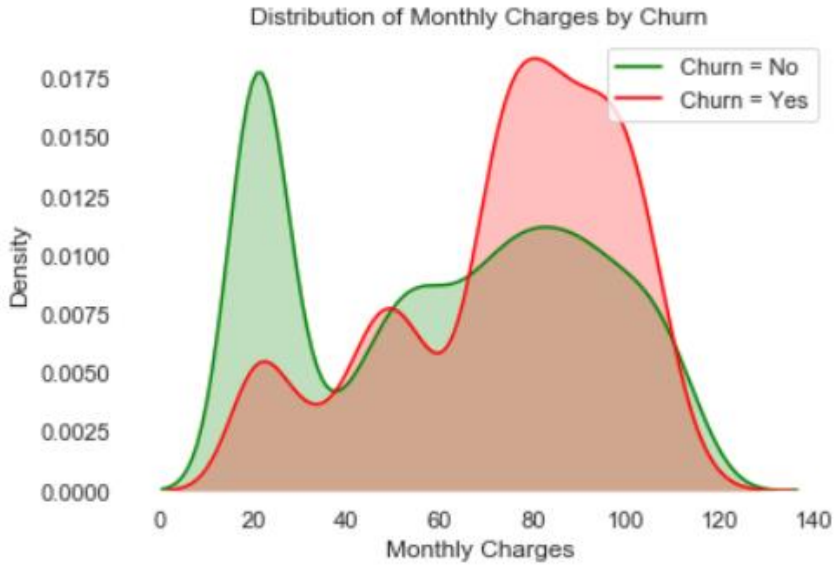
2. Histogram of Tenure of the customers who churned away and who continued showed that the customers with shirter tenure are more likely to churn.



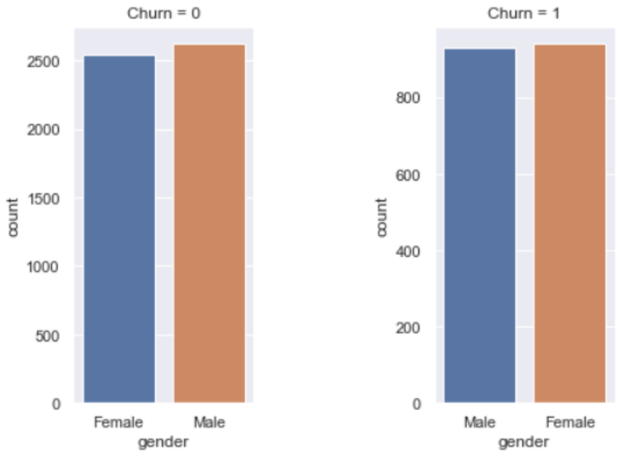
3. Similarly, histogram of Monthly Charges showed that customer who continued the services had lower monthly charges

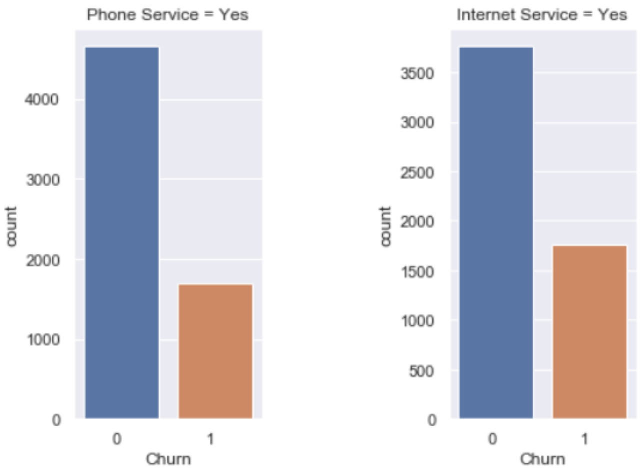


4. We were able to capture the association between churn rates and monthly charges effectively using a Kernel Density Estimation (KDE) plot.

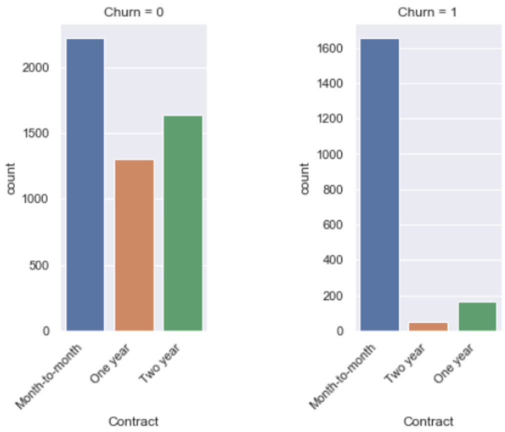


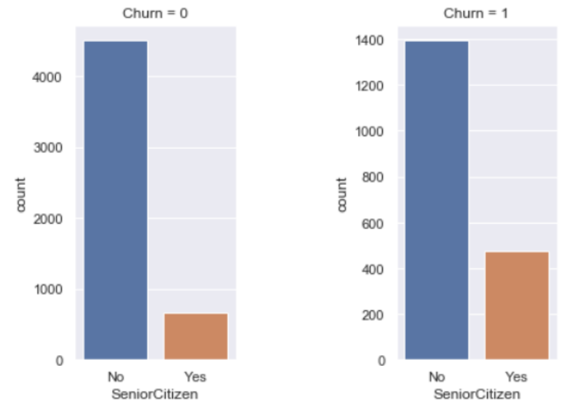
5. Gender and subscription to Phone Service didn’t have a strong association with Churn rates.

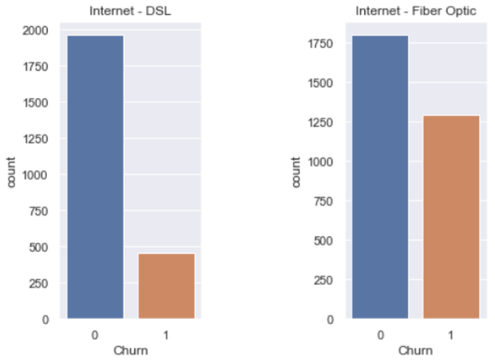


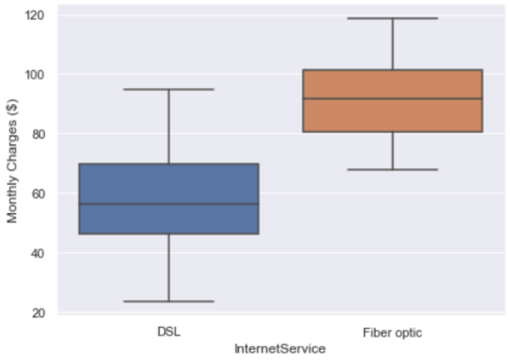


6. On the other hand, seniority, contract-length, and mainly internet service type had strong association with churn rate







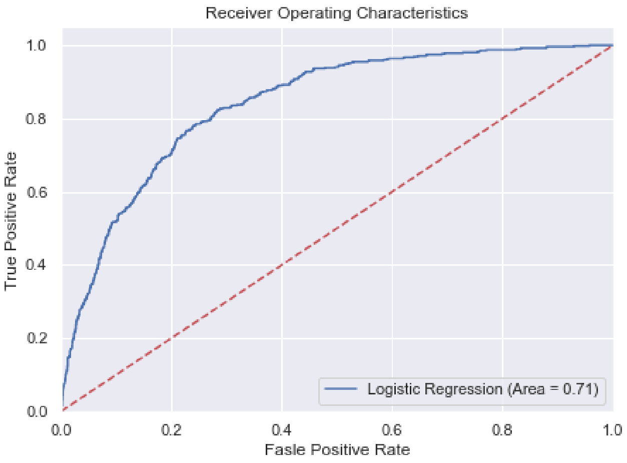


It can be observed in the box plot that the Fiber Optic internet service are more expensive than the DSL internet services

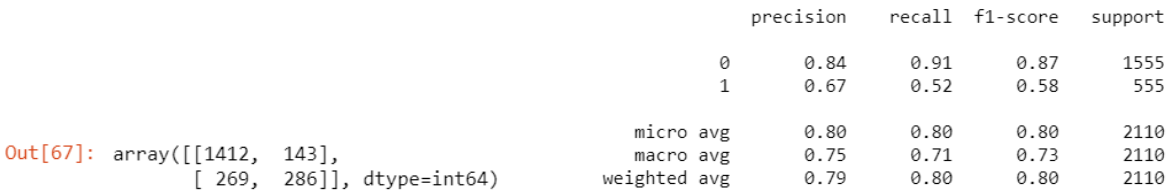
Machine Learning Models and Model Selection

1. Logistic Regression

ROC Curve

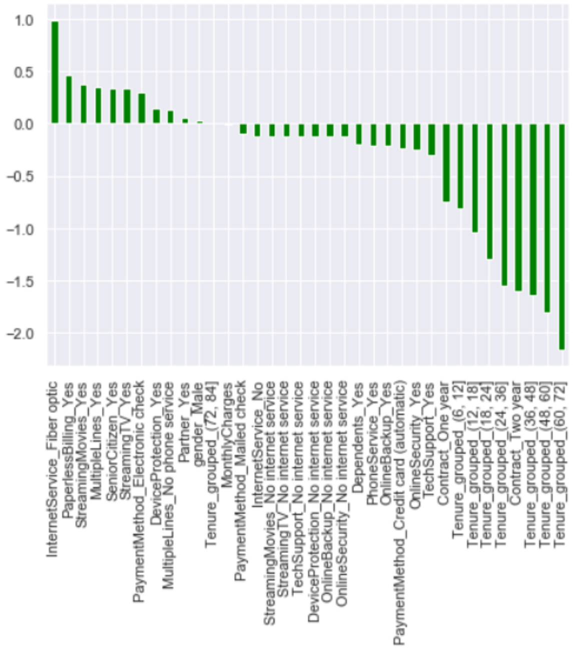


Confusion Matrix and Classification Report



```
                                                          precision    recall  f1-score   support

                                                     0       0.84      0.91      0.87      1555
                                                     1       0.67      0.52      0.58       555

                                             micro avg       0.80      0.80      0.80      2110
Out[67]: array([[1412,  143],                macro avg       0.75      0.71      0.73      2110
                [ 269,  286]], dtype=int64)   weighted avg       0.79      0.80      0.80      2110
```

Variable Importance



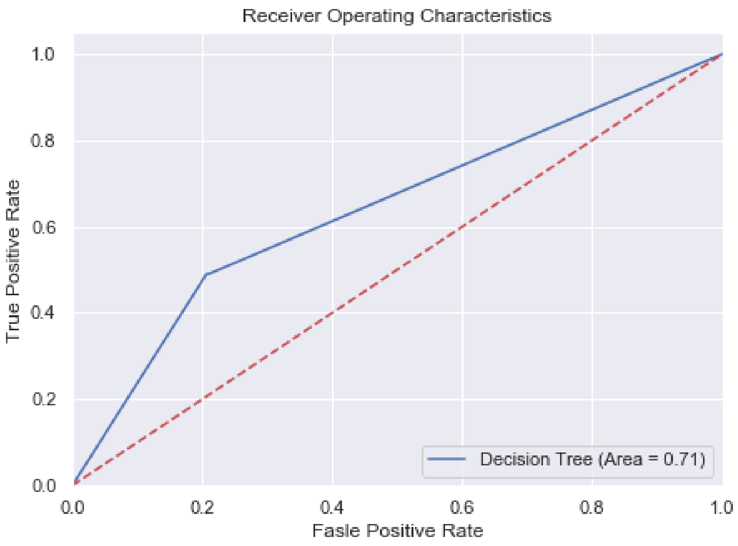
The weights of the variables in the variable importance chart aligns with our exploratory analysis.

- Fiber optic internet service is strongly associated with higher churn rates
- Seniority of the people have a positive relation with churn rates

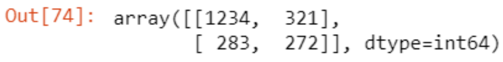
- Longer tenures have a negative relationship with churn rates
- Churn rates diminishes with additional services such as Online backup, Online Security, and Tech support

2. Decision Tree

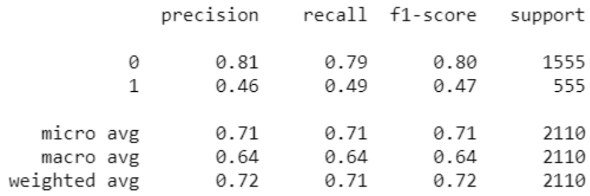
ROC Curve



Confusion Matrix and Classification Report



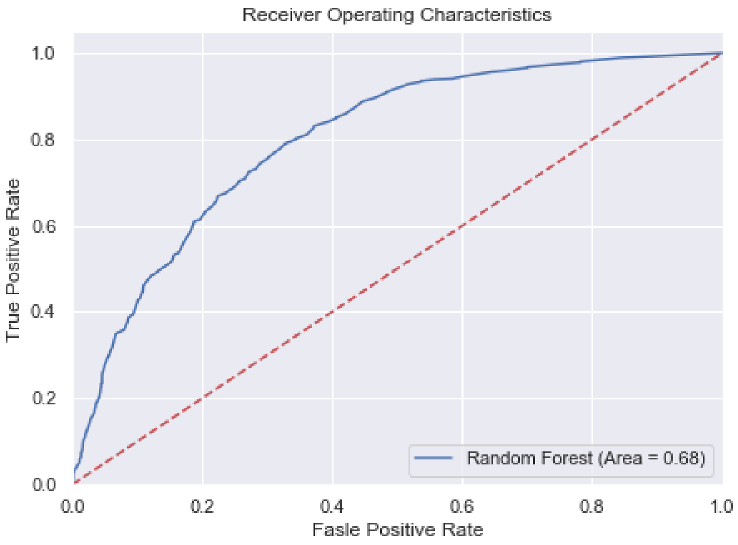
```
Out[74]: array([[1234,  321],
                [ 283,  272]], dtype=int64)
```



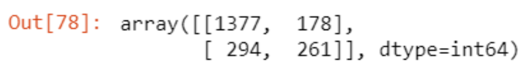
```
             precision    recall  f1-score   support

          0       0.81      0.79      0.80      1555
          1       0.46      0.49      0.47       555

  micro avg       0.71      0.71      0.71      2110
  macro avg       0.64      0.64      0.64      2110
weighted avg      0.72      0.71      0.72      2110
```

3. Random Forest

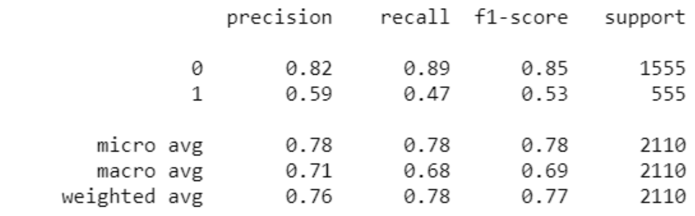
ROC Curve



Confusion Matrix and Classification Report



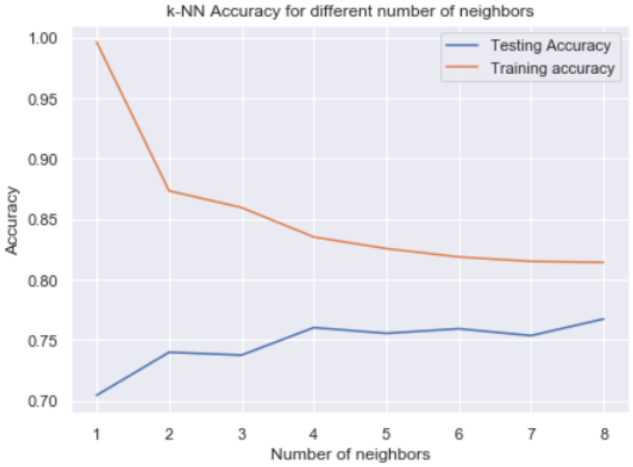
```
Out[78]: array([[1377,  178],
                [ 294,  261]], dtype=int64)
```



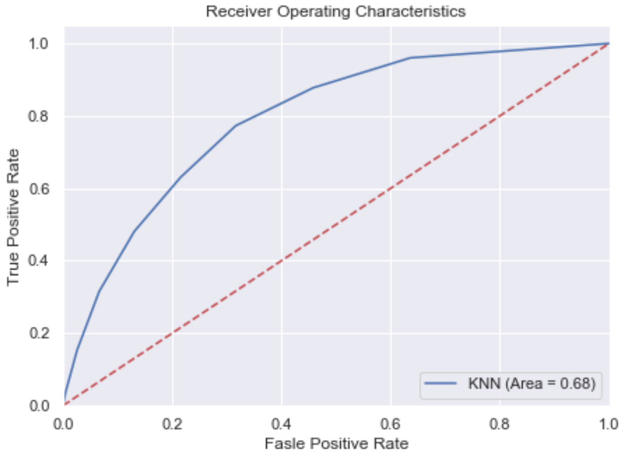
```
             precision    recall  f1-score   support

           0       0.82      0.89      0.85      1555
           1       0.59      0.47      0.53       555

   micro avg       0.78      0.78      0.78      2110
   macro avg       0.71      0.68      0.69      2110
weighted avg       0.76      0.78      0.77      2110
```

4. KNN

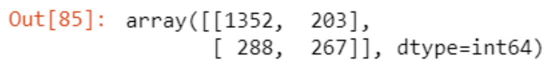
Accuracy for different number of neighbors



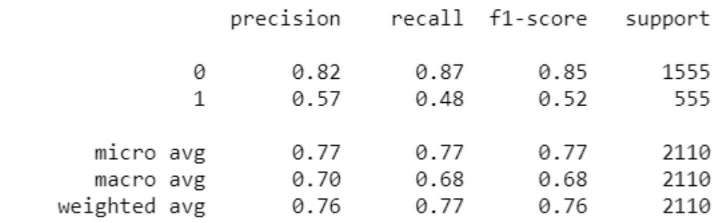
ROC Curve : As the accuracy on test data is the best in case of n=8, we have used n=8 in our model



Confusion Matrix and Classification Report

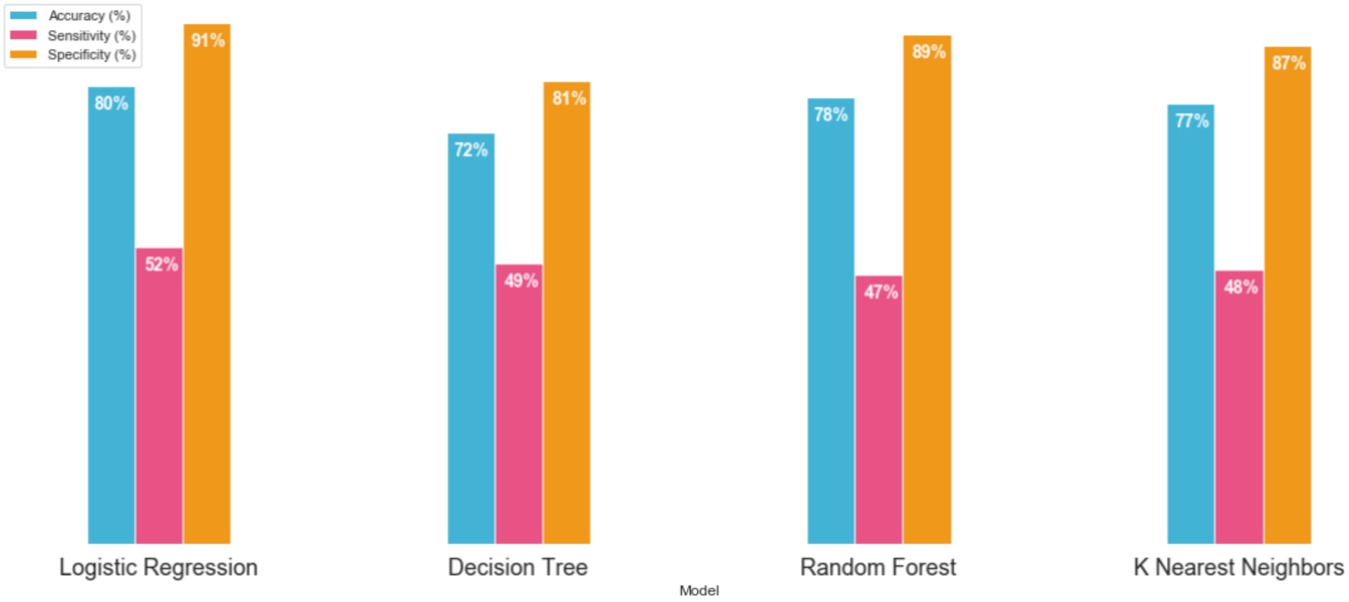


```
Out[85]: array([[1352,  203],
                [ 288,  267]], dtype=int64)
```



```
             precision    recall  f1-score   support

           0       0.82      0.87      0.85      1555
           1       0.57      0.48      0.52       555

   micro avg       0.77      0.77      0.77      2110
   macro avg       0.70      0.68      0.68      2110
weighted avg       0.76      0.77      0.76      2110
```

Comparison of all the models



As mentioned earlier, we will need to focus not only on the accuracy of the model, but also on other metrics such as sensitivity and specificity (whichever is the worst case scenario). Finally, the sensitivity and specificity also depends on the proportion of the negative and the positive instances in the training dataset; model tends to become bias towards the instance (Yes or No / 1 or 0) that occur more frequently – an issue that we will encounter again in our natural language processing exercise.