Data science and machine learning for IBM churn sample data set

Hamidreza Hajiyani ▪ 23.07.2020

Data Analysis:

- Relative number of Churn versus No Churn is 26 to 74 percent

- Relative distribution of each numerical features:

- Low TotalCharges (around 0) and high MonthlyCharges (from 70 to 100) is likely churn.

- Low tenure has higher intensity of churn costumers while at high tenure (more than 50 has a higher) chance of No Churn.

Fig.1 :Ralative distibution of Churn versus No Churn

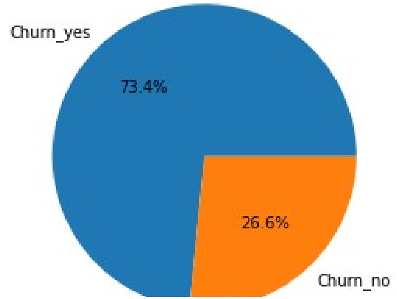


Fig.2a) Distribution of TotalCharges,   Red=Churn

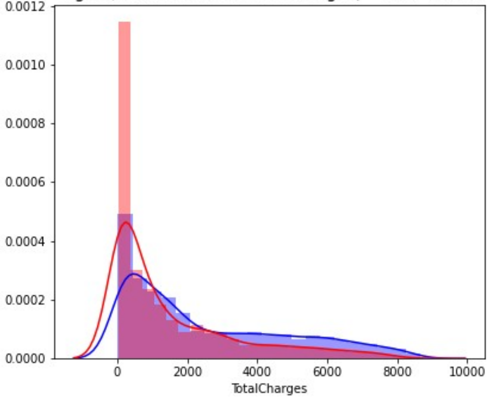


Fig.2b) Distribution of MonthlyCharges,   Red=Churn

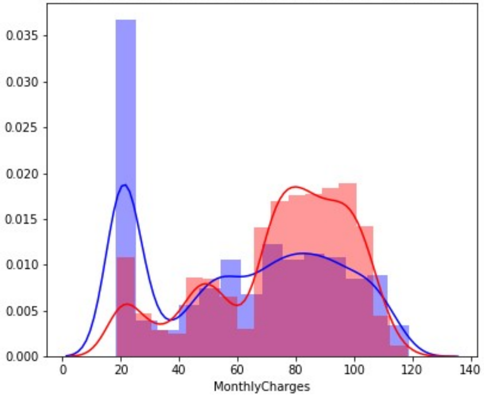
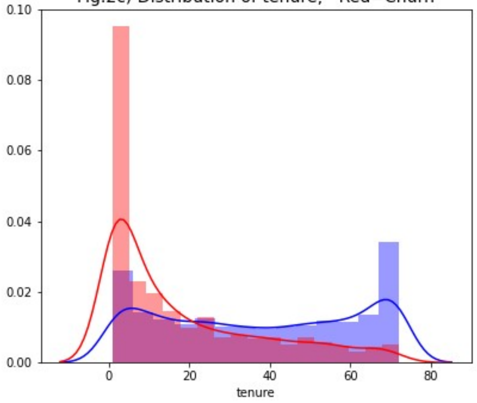
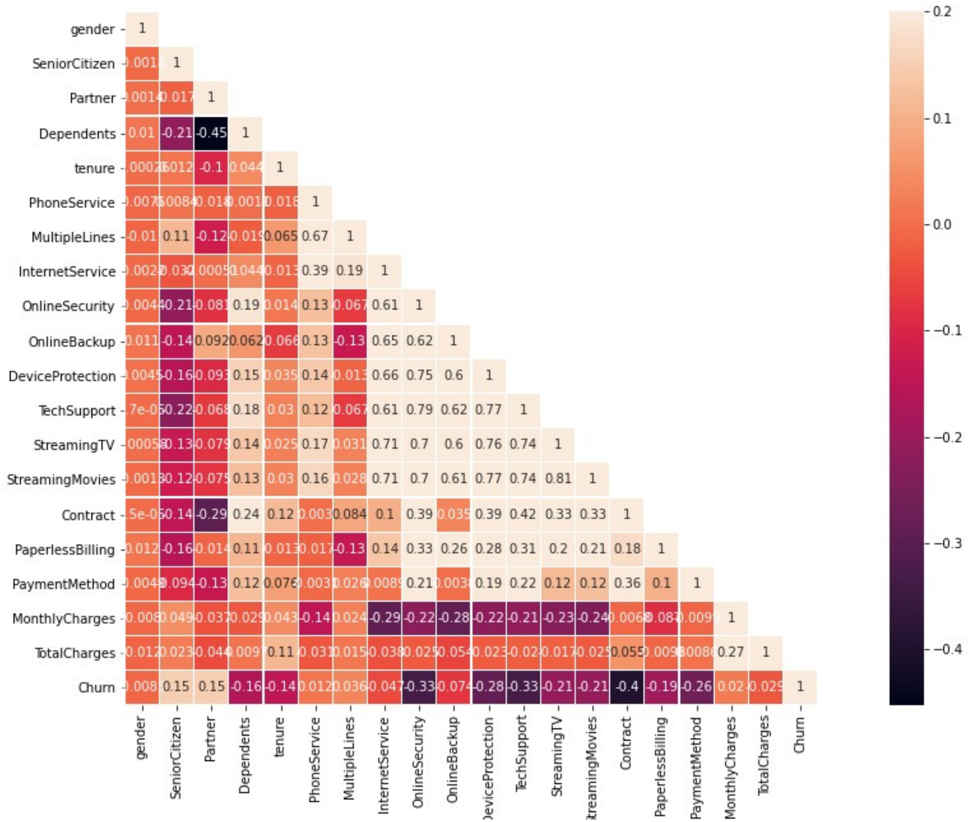
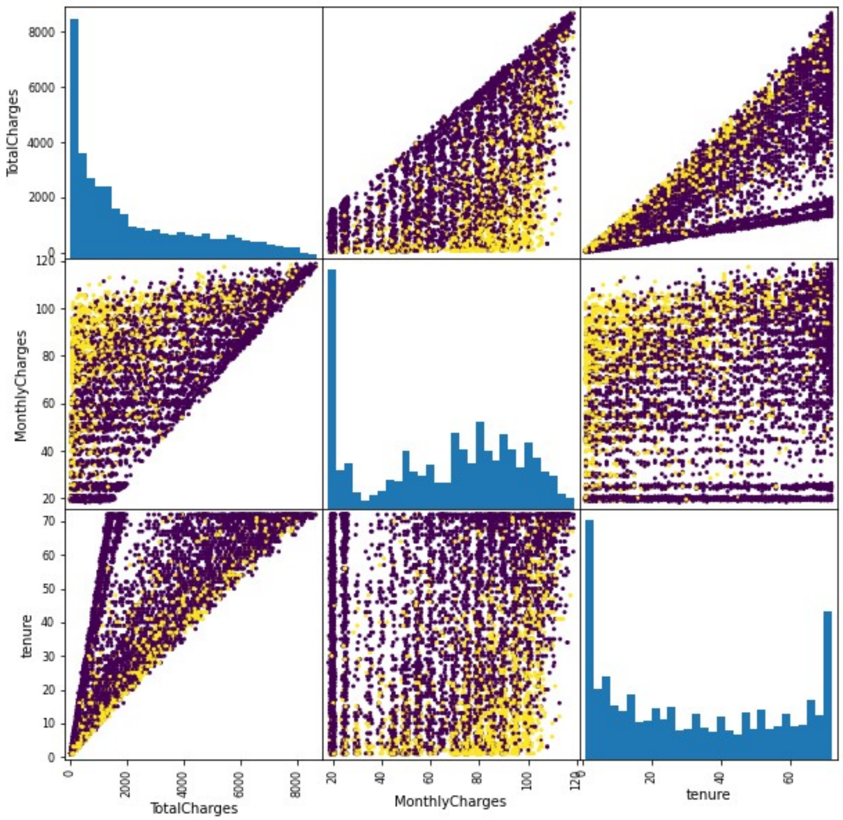


Fig.2c) Distribution of tenure,   Red=Churn



- we illustrate the pair plot of numeral features.
- we can observe boundary and clustering between numerical features.
- Correlation heatmap is also plotted by using of Pearson method





Distribution of categorical features

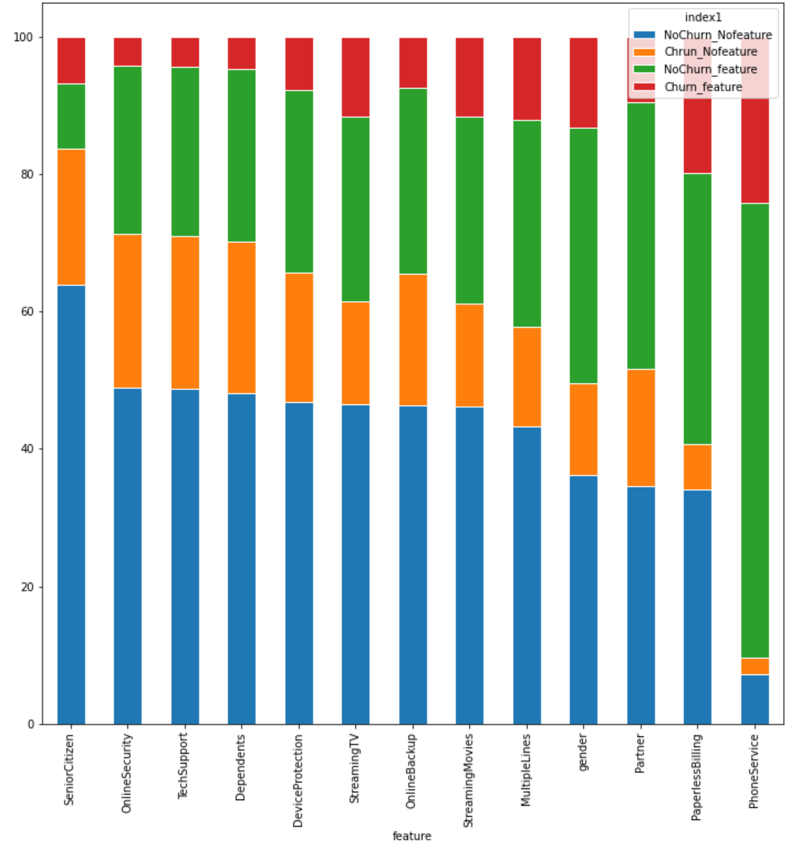
- Distribution of categorical features
To get more insight toward categorical features we calculate the percentage of consumers that are (No)-churn and have that specific feature or not. Of course sum of these groups are 100 percent. Our analysis contain valuable information:

- Only around 15% of the costumers are SeniorCitizens which is the minimum feature. But 40 percent of are churn

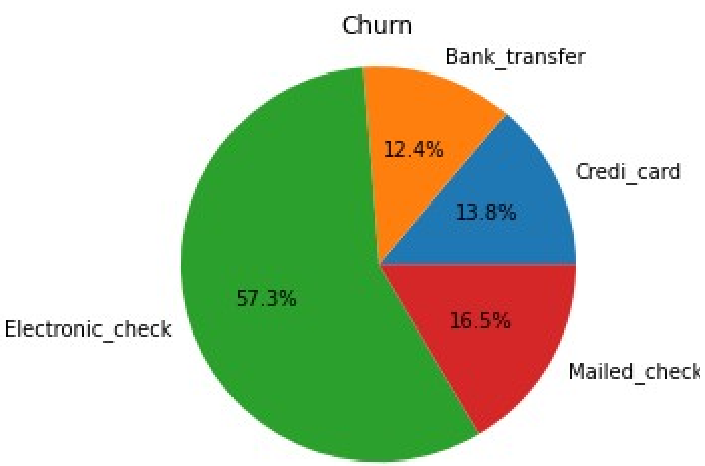
- Only 8% of costumers have no Phoneservice. So we can conclude that is is not a important feature.

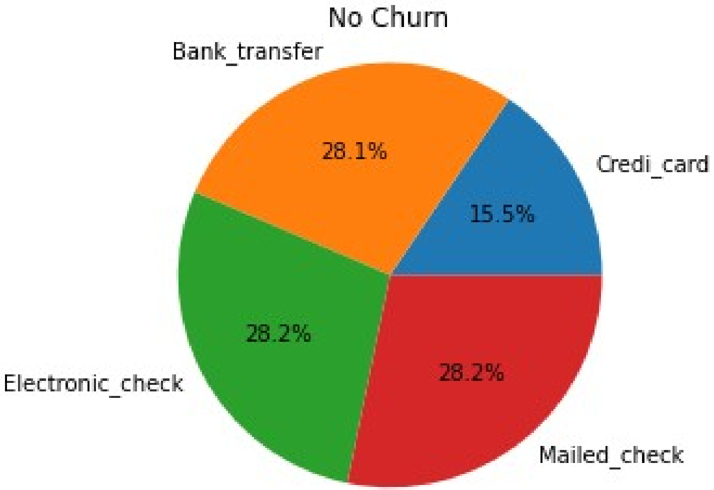
- For the gender also the ration of churn peoples for men and women are both around 50 percent.

- For the partner, around 1/2 of people with no partner are churn. But around 1/3 for people with partner are churn.



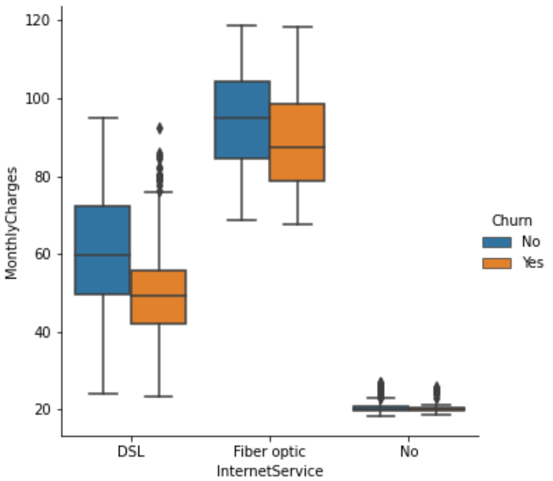
Distribution of categorical features: PaymentMethod

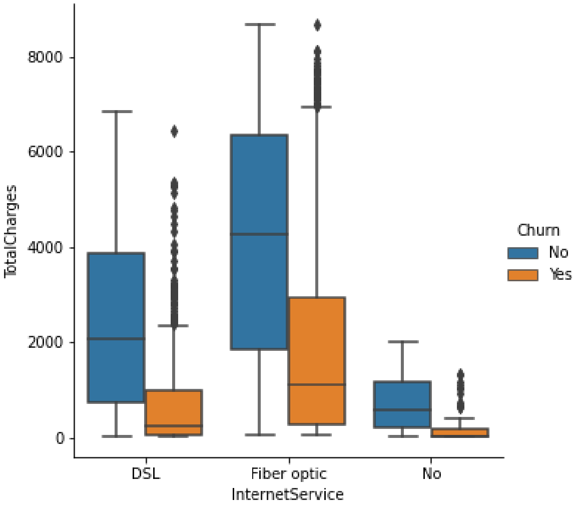


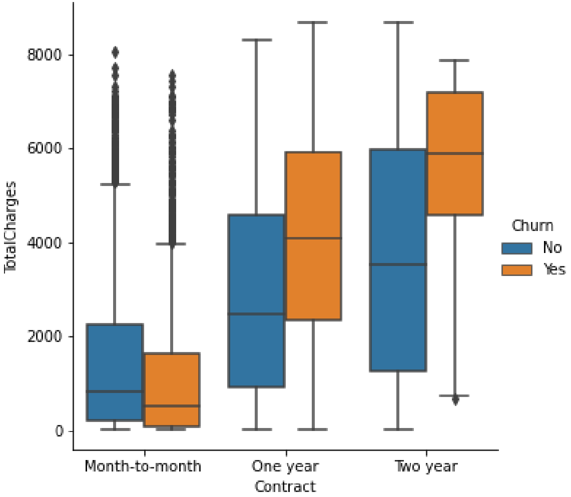


- The pieplot for the PaymentMethod indicates that more that half of churn people use the electronic check at most and Mailed check at least. However, the No churn people use different methods equally.

Correlation of numerical and categorical features







1. InternetService versus MonthlyCharges

- People that use the DSL with MonthlyCharges more than 60 are more likely to be No Churn.
- For the Fiber optic InternetService this number should be more than 100.

2. InternetService versus Totalchages

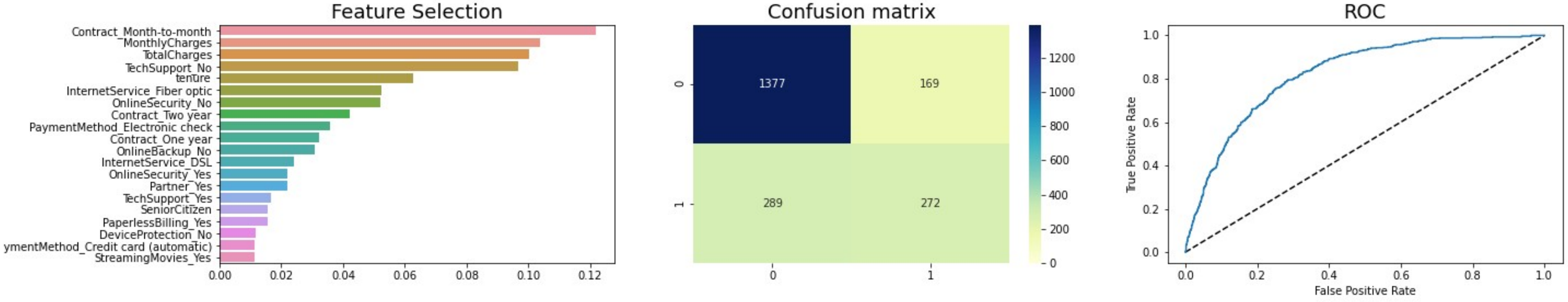
- For DSL users when the TotalCharges are higher that 1800, the No Chrun is more likely, however, for Fiber optic InternetService this number is around 4000.

3. Contract versus Totalchages

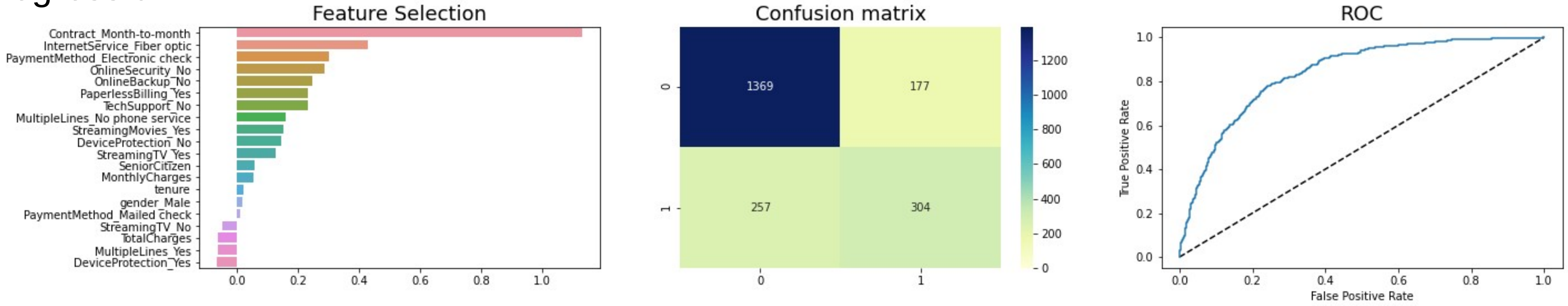
- Short term contracts has higher chance to be No churn but when the contract is longer it has higher chance to be churn.

Machine learning methods:

GradientBoostingClassifier



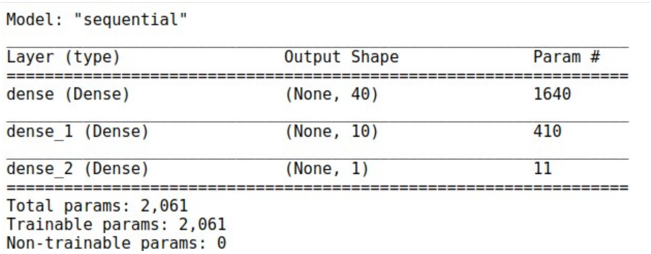
LogisticRegression



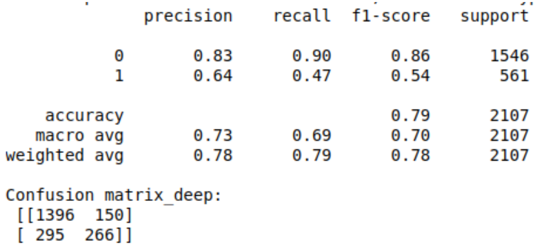
While both methods provide satisfying results Gradient boosting provide better feature selections

Machine learning methods: Deep learning

For the deep learning we have used the sequential method and one hidden layer. Different number of hidden layers and other hyper-parameters were tested but we did not get any significant improvements.



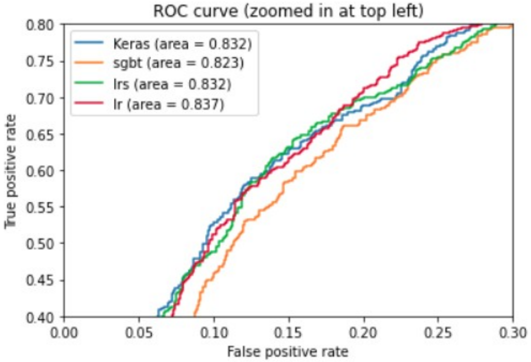
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 40)                1640
_________________________________________________________________
dense_1 (Dense)              (None, 10)                410
_________________________________________________________________
dense_2 (Dense)              (None, 1)                 11
=================================================================
Total params: 2,061
Trainable params: 2,061
Non-trainable params: 0
```

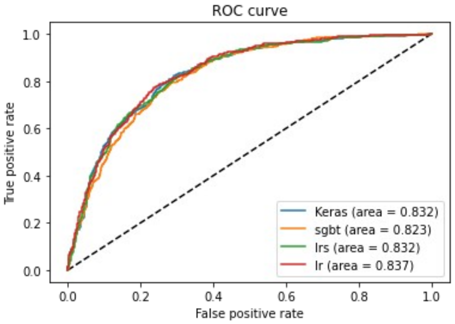


```
              precision    recall  f1-score   support

           0       0.83      0.90      0.86      1546
           1       0.64      0.47      0.54       561

    accuracy                           0.79      2107
   macro avg       0.73      0.69      0.70      2107
weighted avg       0.78      0.79      0.78      2107

Confusion matrix_deep:
 [[1396  150]
 [ 295  266]]
```

Conclusion:
Here we can see the comparison of different models. They show almost similar trend.

These results have been concluded after a short time of research. More precise results can be achieved by hyper-patameters tuning or adopting of different machine learning models.


ROC curve (zoomed in at top left)
Keras (area = 0.832)
sgbt (area = 0.823)
lrs (area = 0.832)
lr (area = 0.837)
True positive rate
False positive rate


ROC curve
Keras (area = 0.832)
sgbt (area = 0.823)
lrs (area = 0.832)
lr (area = 0.837)
True positive rate
False positive rate