Data science and machine learning for IBM churn sample data set

Hamidreza Hajiyani ▪ 23.07.2020

Data Analysis:

- Relative number of Churn:1 versus Churn:0 is 26 to 74 percent, respectively.
- 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

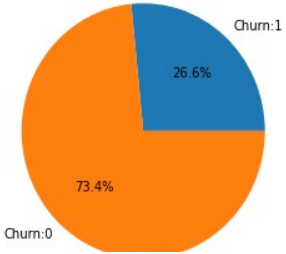


Fig.2a) Distribution of TotalCharges, Red=Churn:1

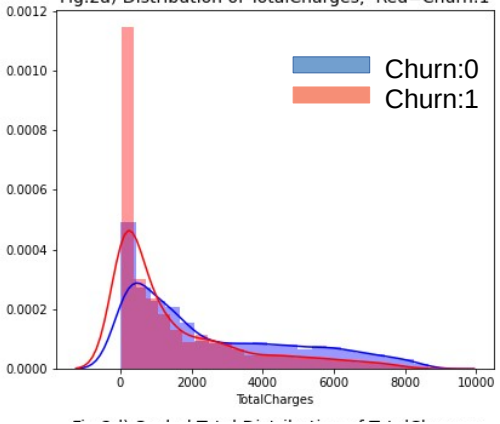


Fig.2b) Distribution of MonthlyCharges, Red=Churn:1

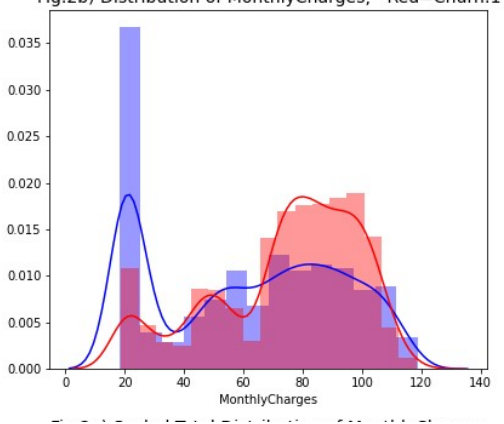
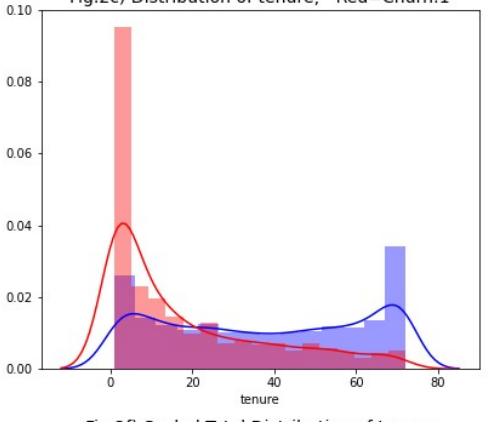
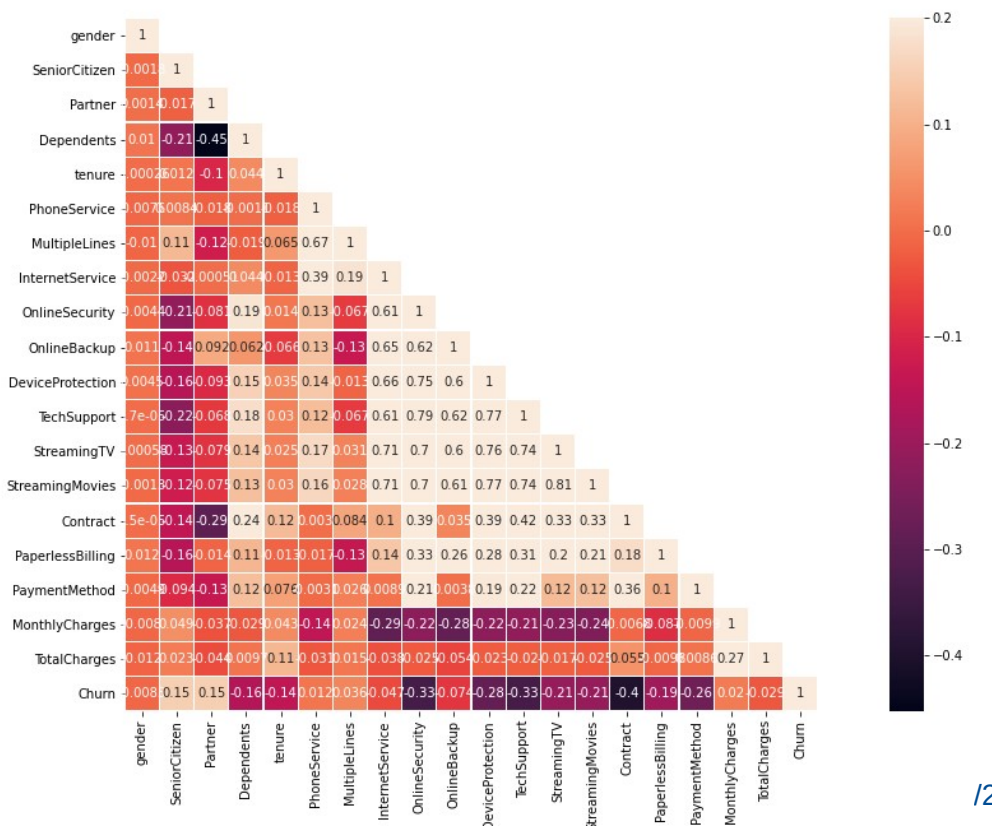
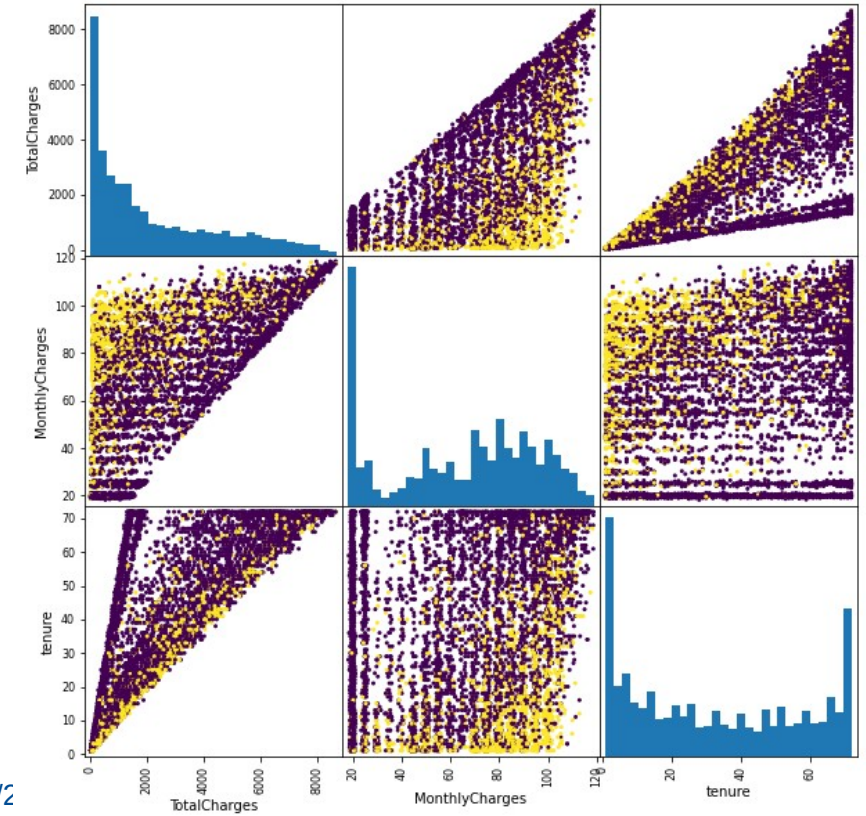


Fig.2c) Distribution of tenure, Red=Churn:1



- 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

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

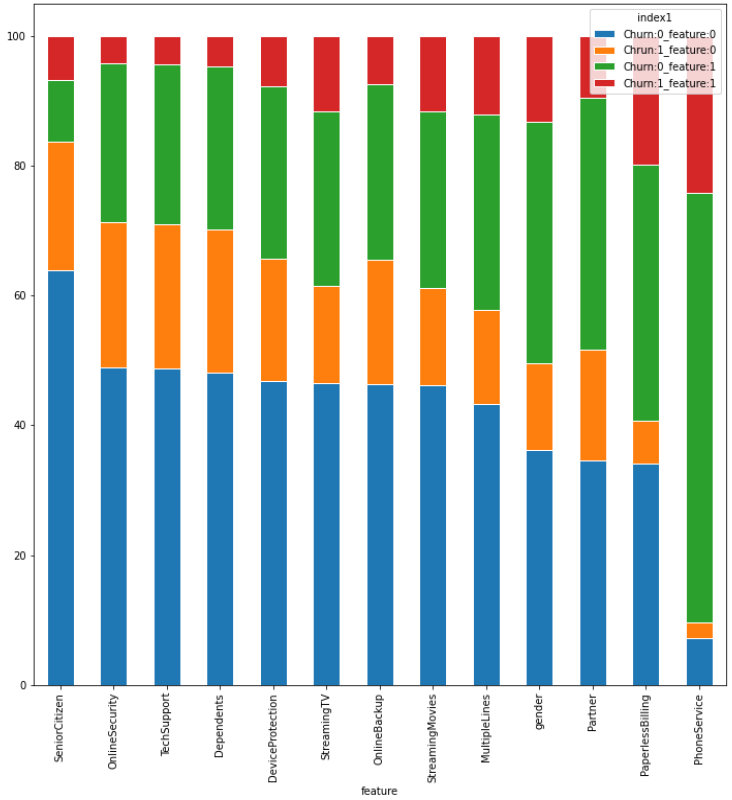
- Distribution of categorical features
To get more insight toward categorical features we calculate the percentage of consumers that are (No)-churn and have that specific feature or not. Of course sum of these groups are 100 percent. Our analysis contain valuable information:

- Only 17% of the costumers are SeniorCitizens which is the minimum feature. But 40 percent of them are churn (Ratio of red and green)

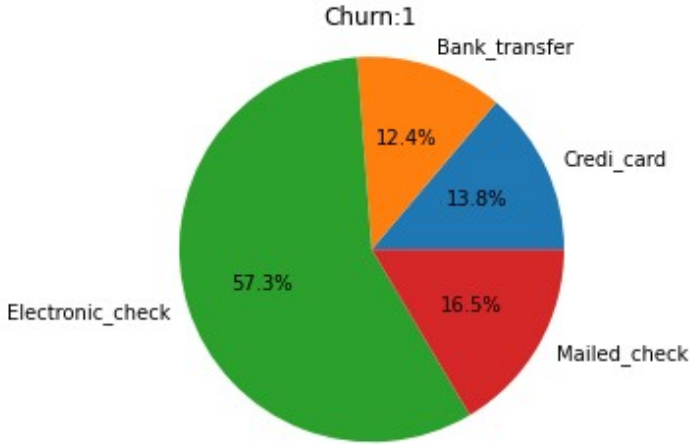
- Only 8% of costumers have no Phoneservice. So we can conclude that is is not a important feature.

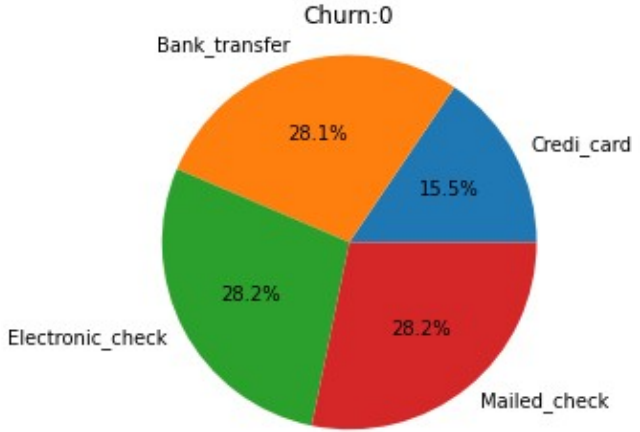
- For the gender also the ratio of churn peoples for men and women are both around 50 percent (Ratio of red and orange).

- For the partner, around 1/3 of people with no partner are churn (orange). But around 1/4 for people with partner are churn (red).



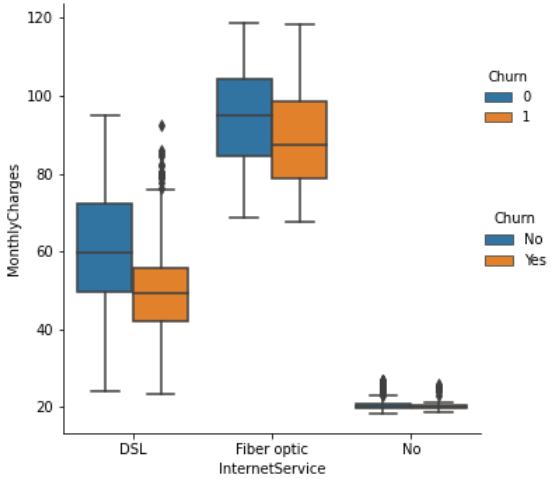
Distribution of categorical features: PaymentMethod

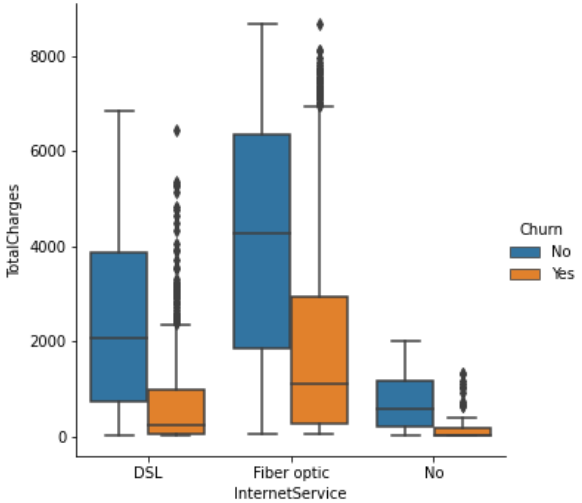


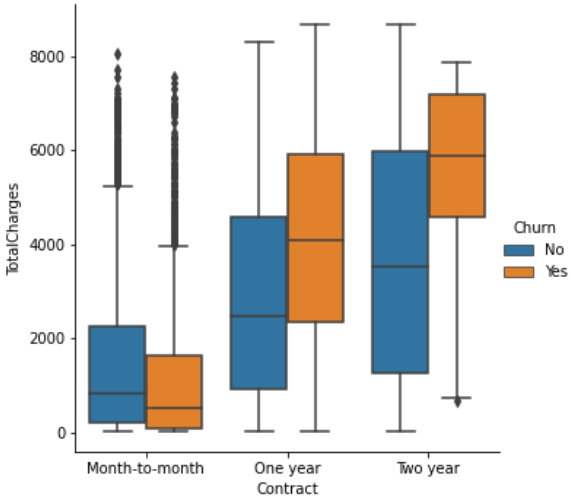


- The pieplot for the PaymentMethod indicates that more than half of churn people use the electronic check at most and Mailed check at least. However, the No churn people use different methods equally.

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 Churn 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
- Sequential correction of predecessor's errors.
- Does not tweak the weights oft raining instances.
- Fit each predictor is trained using its predecessor's residual errors as labels.
- Since in each predictor level we have same number of rows the split point of many tree might be same..

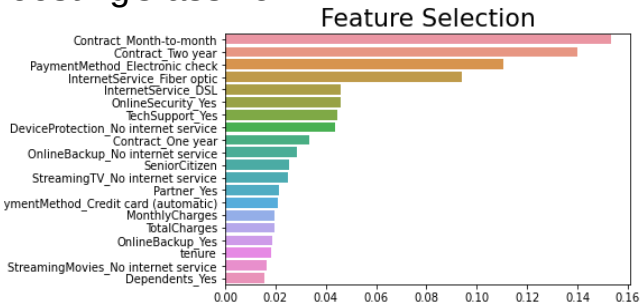
Stochastic Gradient Boosting
- Fraction of features (X_train) should included for each predictor, flag: max_features
- Ratio of including data flag: subsample
- Result: further ensemble diversity.
- Effect: adding further variance to the ensemble of trees

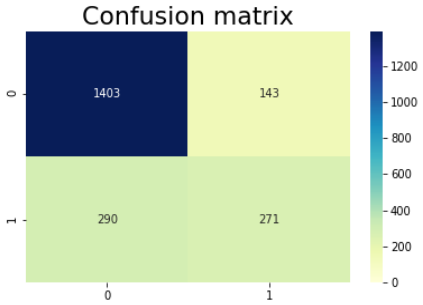
Support Vector Machines
- The classifier separates data points using a hyperplane with the largest amount of margin.
- SVM uses a technique called the kernel trick that takes a low-dimensional input space and transforms it into a higher dimensional space.

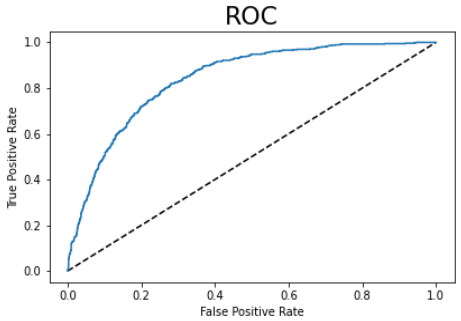
Machine learning methods:

UNIVERSITÄT DUISBURG ESSEN
Open-Minded

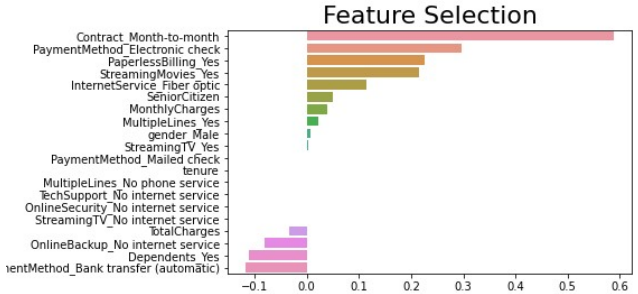
GradientBoostingClassifier

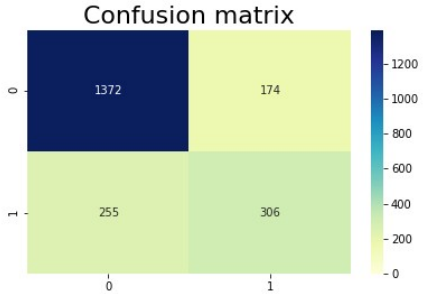


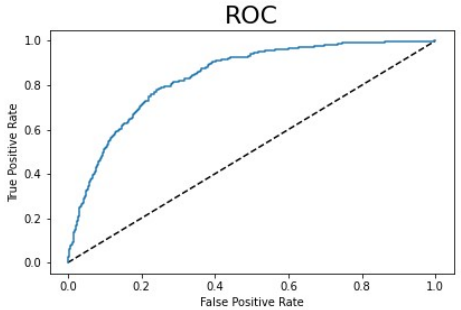




LogisticRegression



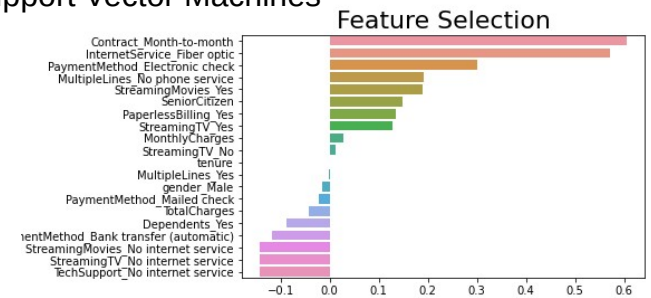


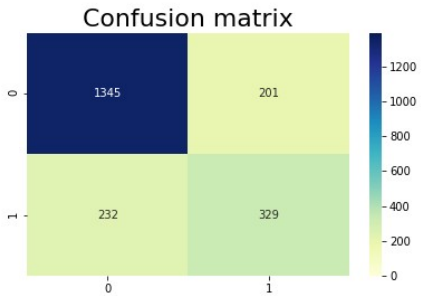


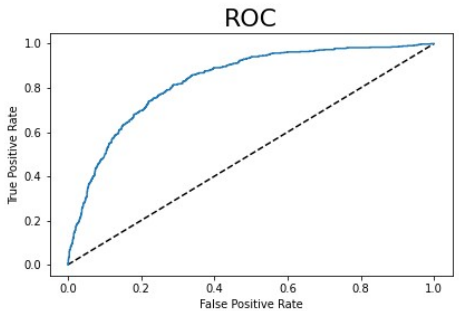
While both methods provide satisfying results Gradient boosting provide better feature selections

Machine learning methods

Support Vector Machines







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.

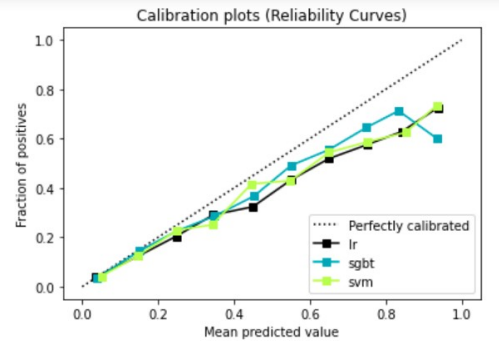
```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 40)                1640
_________________________________________________________________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

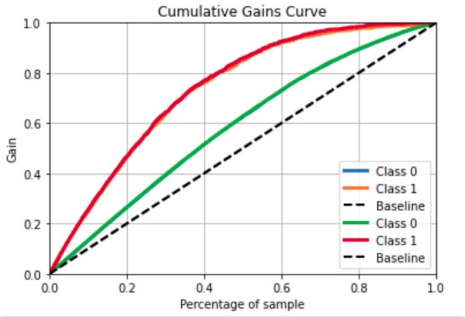
	precision	recall	f1-score	support
0	0.85	0.88	0.86	1546
1	0.63	0.57	0.59	561
accuracy			0.79	2107
macro avg	0.74	0.72	0.73	2107
weighted avg	0.79	0.79	0.79	2107

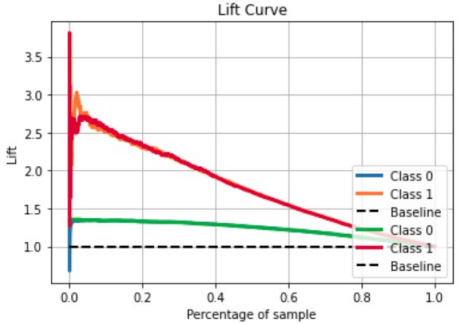
```
Confusion matrix_deep:
 [[1356  190]
 [ 243  318]]
```

Evaluation measures

Comparison of
1. Calibrations
2. Cumulative Gains
3. Lift Curve







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 or uplift modeling in real-world business