churn-draft-2

Introduction to Churn

Customer churn is one of the main issues in fields such as telecommunications, internet service providers, e-commerce, marketing, or banking. In the telecomunication industry, the market is saturated, making it very difficult to attract new customers. On the customer side, it is very easy to switch to a new provider if it provides a more beneficial offer (Canale and Lunardon, 2014). Because of this saturated market, the cost of aquiring new customers can be five to six times higher than retaining existing customers. Therefore, companies might be best invested in developing customers' trust on the service rather than attracting new customers.

Researches on predicting customer churn have been done to try to predict whether a customer is likely to quit the service of their service provider and join a different one. Having an understanding of the influential factors causing customer churns on the service can help companies understand customers' needs and adjust on their provided services based on these factors to reduce churn.

A lot of classification methods have been used to predict the rate of churn. One of those is logistic regression.

The hypothesis for the following project is that given the same data set, all methods if accurated performed should yield similar results. All of these models will have the same question, which is, what predictive models perform the best among the models used in this project to predict the churn rate in the telecom industry. I hypothesize that complex models do not necessarily perform better than simplier models.

The description of the data set

Our data set consists of 7043 profiles of telecom customers and is available via.... The data contains of two main types of variables:

- Customers' personal characteristis such as their gender, whether they have a partner, their tenure status, whether they are senior citizens.
- There usage behaviors such as phone service, internet service, online security, tech support, streaming TV, streaming movies, contract, payment method and their month charges.

Both of these types of variables are beneficial for our predictive analysis. It helps us to detect whether their decisions to churn is based on personal characteristics or on their service usage behaviors.

Method

This project uses R to implement three predictive models: logistic regression, random forest and decision trees. I choose these models because they are often used in papers that have been published on churn. These are also well-known predictive models.

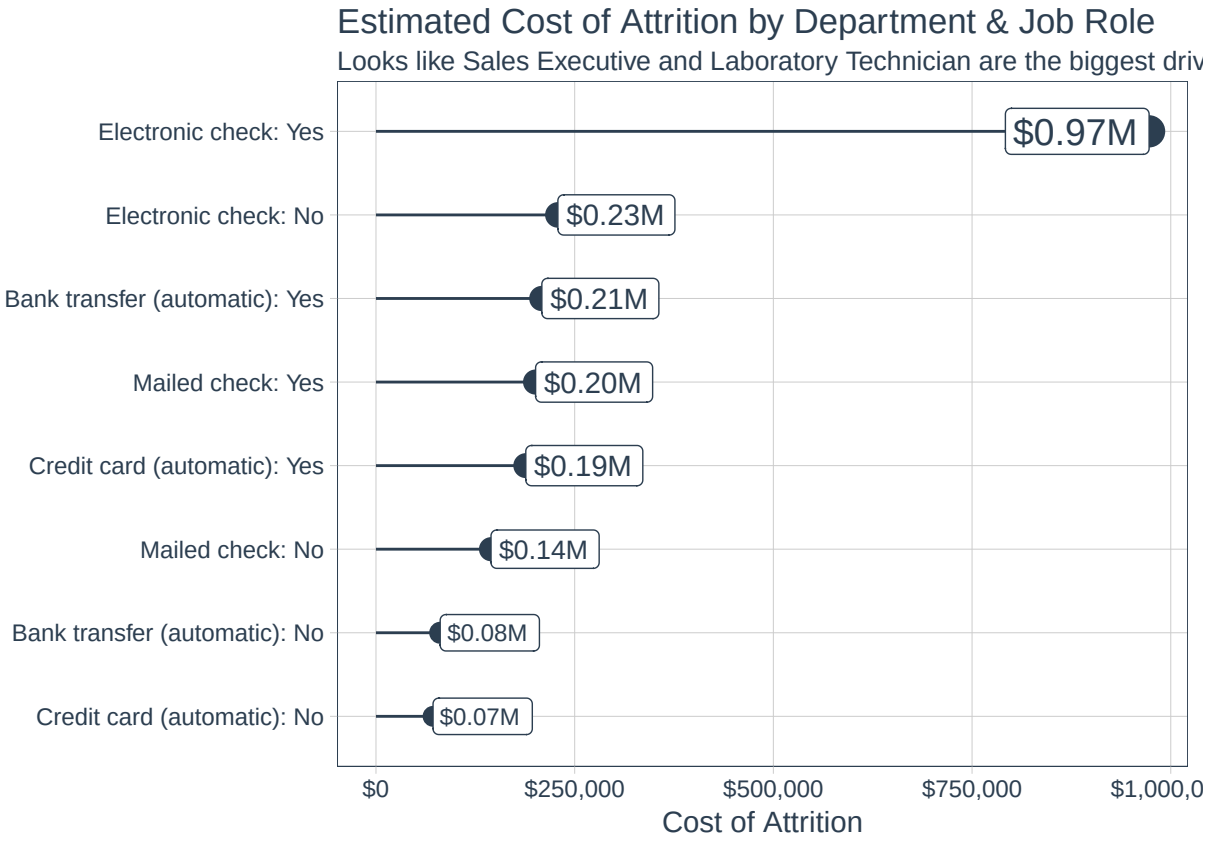
Some of them will be implemented as black box, meaning I cannot know very well what are the processes they do the analysis on. Therefore, I will choose the dive deep into logistic regression and see if they are well performed. I choose logistic regression particularly because it is one of the models easiest to explain and measure performance. Bear in mind that these models will be briefly explained instead of looking into details. Specifically, I will mention its advantages and disadvantages of each of these models, then I will explain the steps I did to get the final results.

Business Understanding

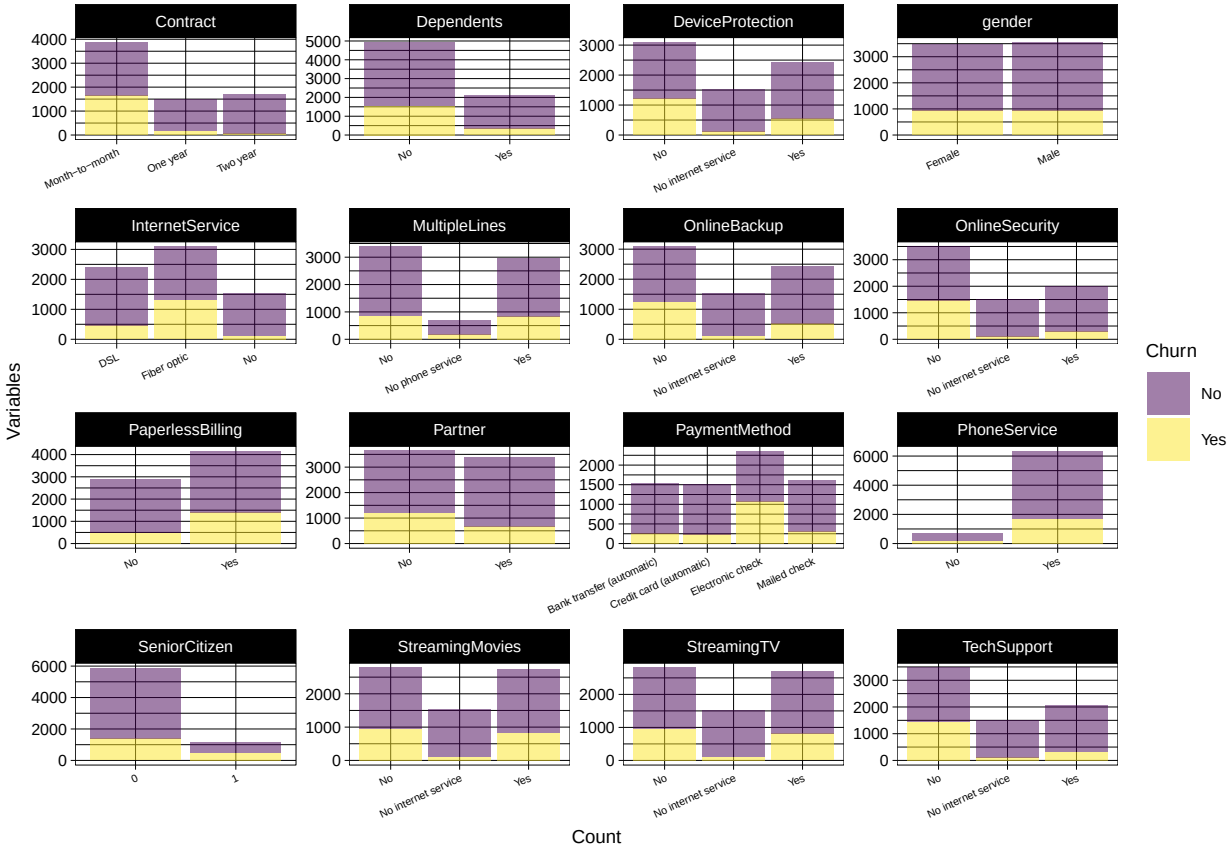
summary table

	Churn	n	pct
1	No	5174	0.73
2	Yes	1869	0.27

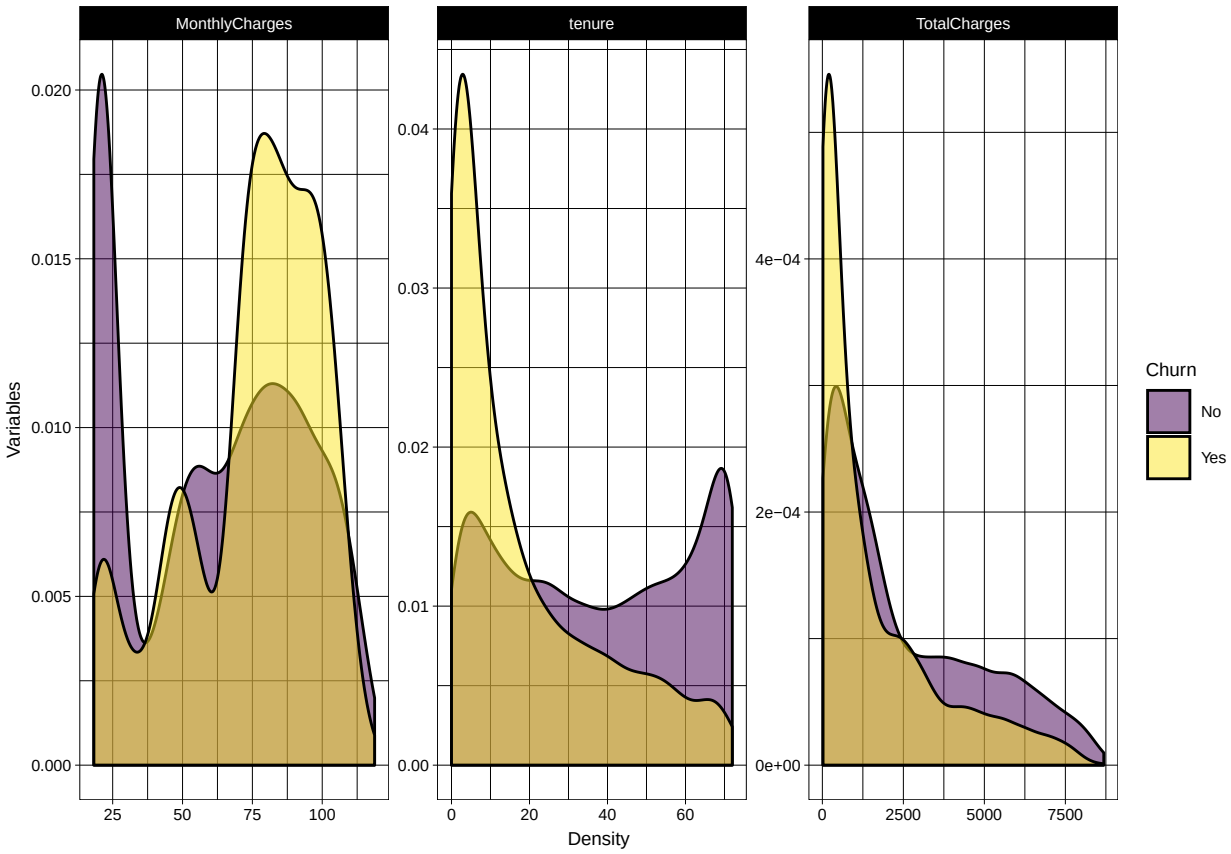
code



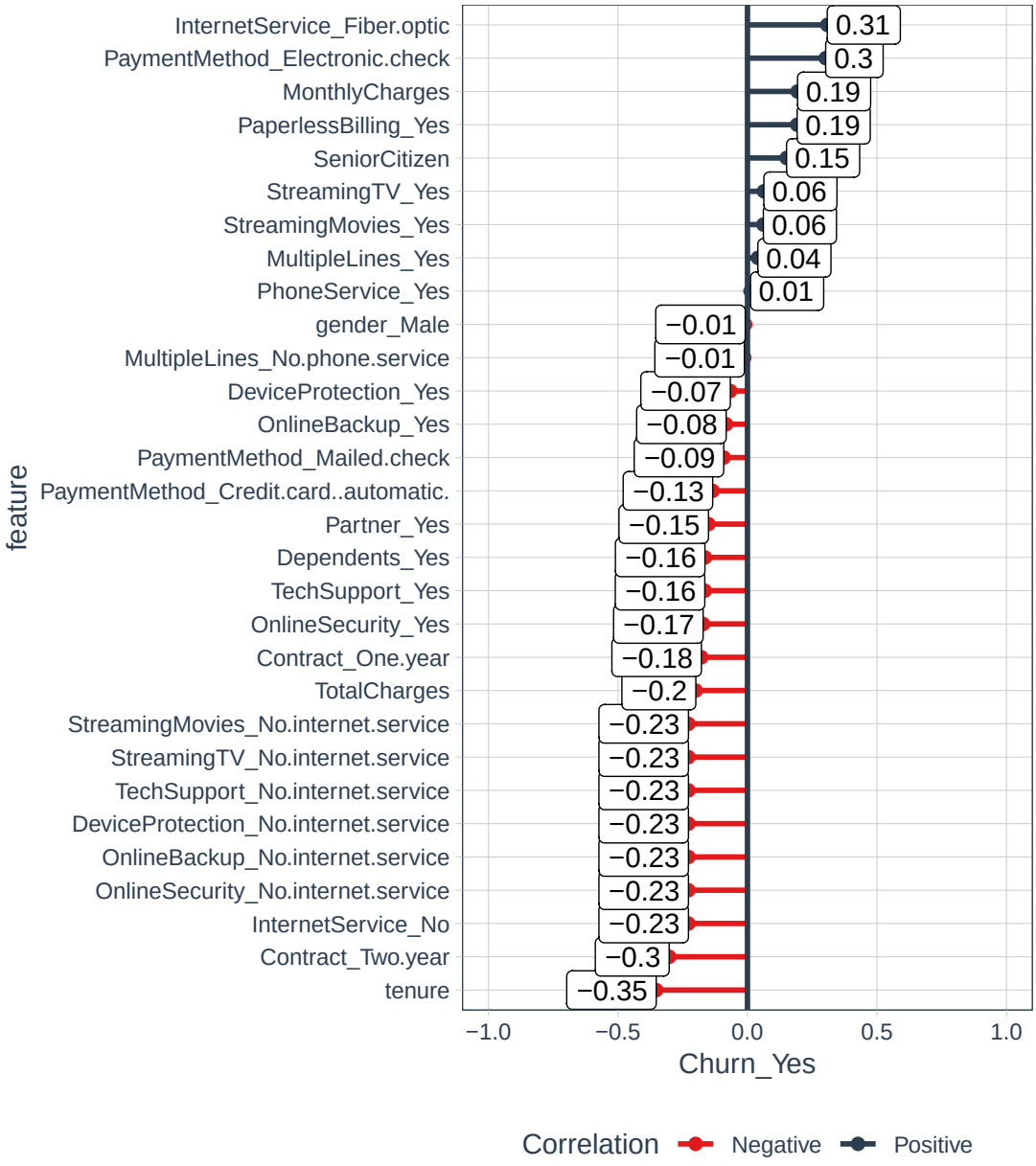
Data Understanding



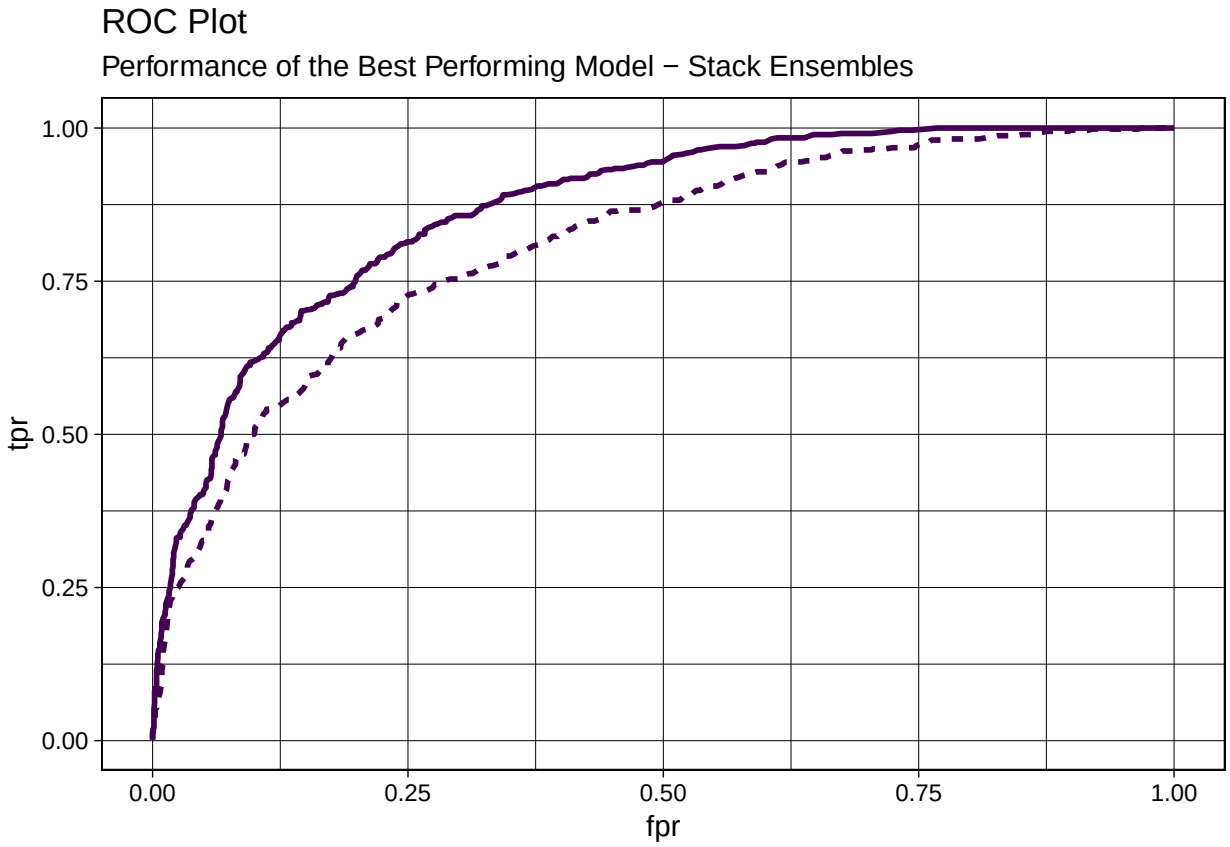

```
## Warning: Removed 11 rows containing non-finite values (stat_density).
```

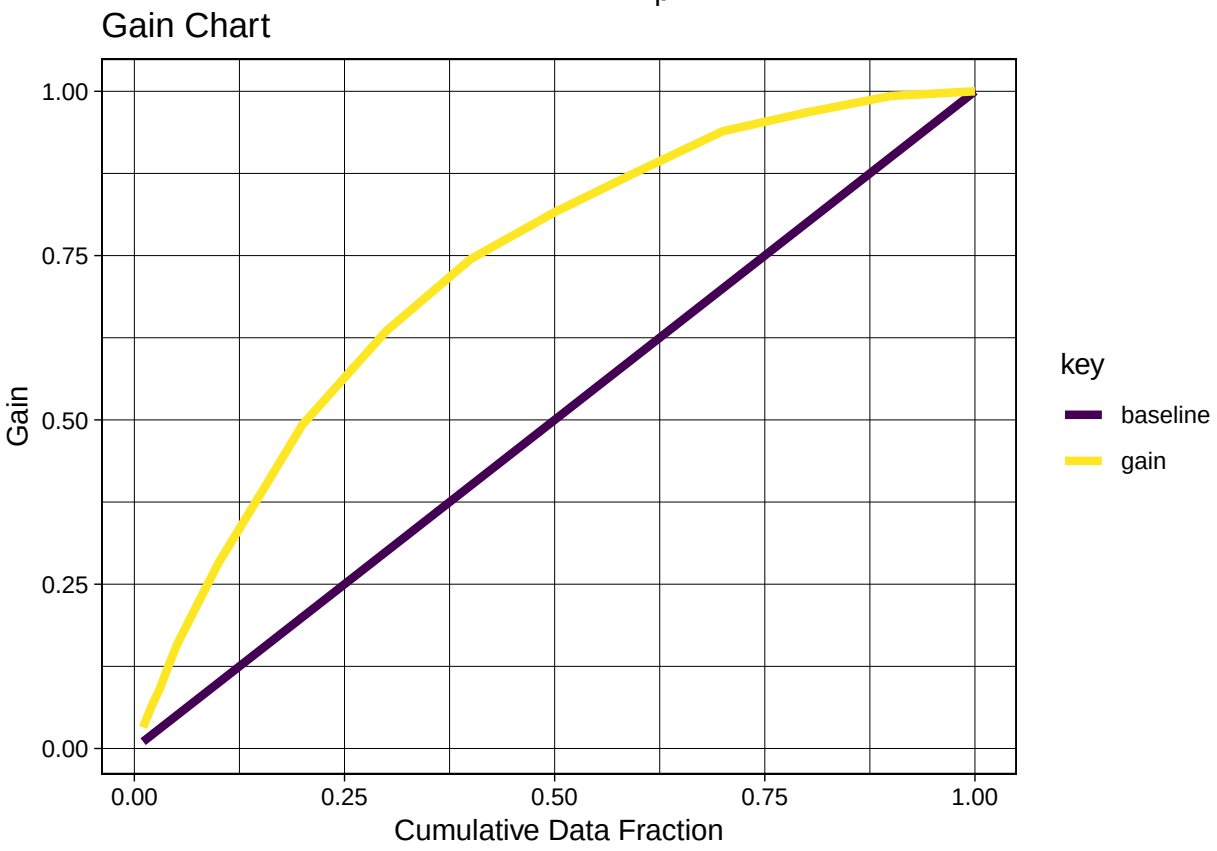

MonthlyCharges
tenure
TotalCharges
Variables
Density
Churn
No
Yes
0.020
0.015
0.010
0.005
0.000
0.04
0.03
0.02
0.01
0.00
4e−04
2e−04
0e+00
25
50
75
100
0
20
40
60
0
2500
5000
7500

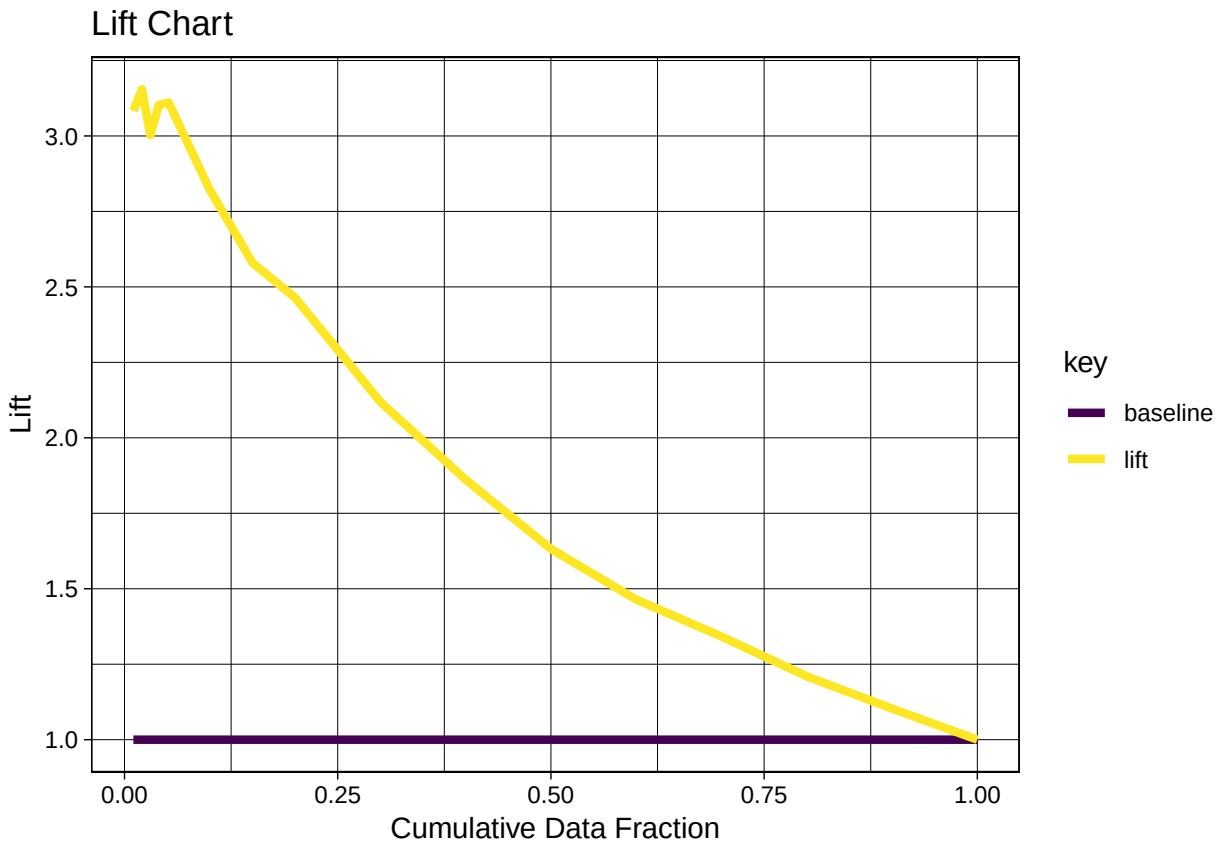
Correlation Analysis



Modeling


ROC Plot
Performance of the Best Performing Model − Stack Ensembles
tpr
fpr
0.00
0.25
0.50
0.75
1.00


Gain Chart
Gain
Cumulative Data Fraction
key
baseline
gain
0.00
0.25
0.50
0.75
1.00


Lift Chart
Lift
3.0
2.5
2.0
1.5
1.0
0.00
0.25
0.50
0.75
1.00
Cumulative Data Fraction
key
baseline
lift