Stockify Final Regression Report

Abdullah Salau

December 2021

1 Introduction

This is a brief report of the final scores and accuracies of Stockify's Regression models. It summarizes details about the dataset, the preprocessing techniques involved, and the experiments undergone by the different models attempting to predict stock performance.

2 Preprocessing

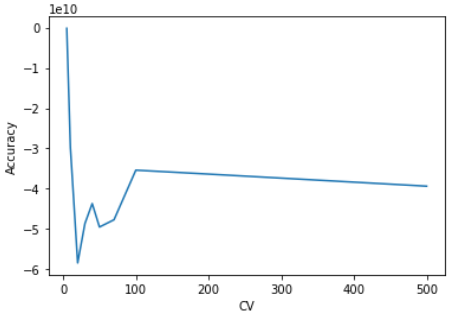
The Stock Dataset consists of 225 features, including 222 predictors, 1 id column, 1 output variable for classification, and another output variable for regression. The dataset also contains 3808 records. To clean the dataset, Stockify deletes the ID column and the output attribute for classification, and then correctly labels the predictors as the inputDF, and the output attribute for regression as the outputSeries. The predictors of the featured dataset also contain numerous missing values which are then filled with either that attribute's mean or median, depending on the attribute's skew. Additionally, categorical variables are taken care of using one-hot encoding.

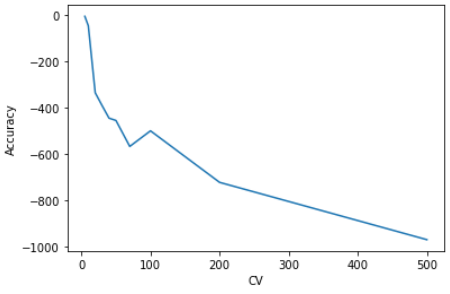
For better model performance, Stockify utilizes advanced preprocessing techniques such as removing attributes with low variation, removing attributes with low correlation with the target variable, log transformations, Hash encoding, isolation forests, etc.

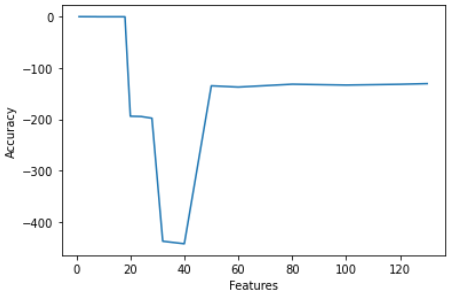
3 Model Performance(Using 'r2" scoring)

3.1 Multiple Linear Regression

- Initial Cross Validation results with standardization to regularize dataset
- Removing variables with low correlation with target variables (x¡—0.1—) leaves only 5 features which give the following results for different CV folds:
- Using selectKBest Features:

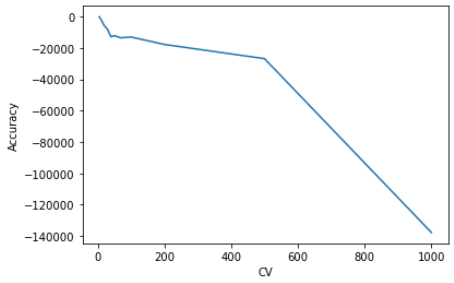

1e10
0
−1
−2
−3
−4
−5
−6
Accuracy
0
100
200
300
400
500
CV


0
−200
−400
−600
−800
−1000
Accuracy
0
100
200
300
400
500
CV

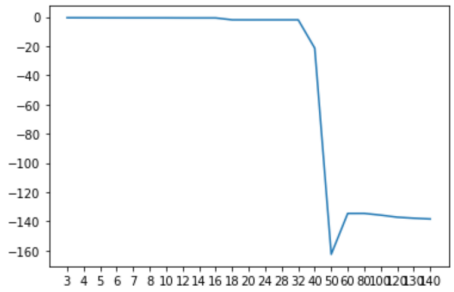

0
−100
−200
−300
−400
Accuracy
0
20
40
60
80
100
120
Features

3.2 Ridge Regression

- Initial Cross Validation results with Normalization to regularize dataset



- Using selectKBest, with removing attributes with low correlation with target variable but reducing threshold to increase features (—x—¡0.001 gave us 143 features instead of 15), 5CV folds, and normalize2 (r2 scoring)



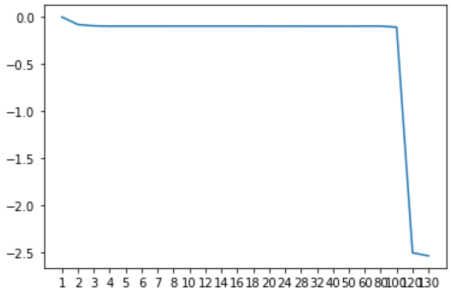
- Using Grid Search for hyper parameter tuning gives us the following setting for our Ridge model: Best Hyperparameters: 'alpha': 100, '$fit_intercept'$: $False,'$ $normalize'$: $True$

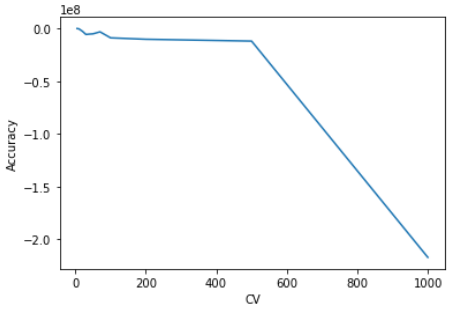
- SelectKBest features with grid Search hyperparameter settings

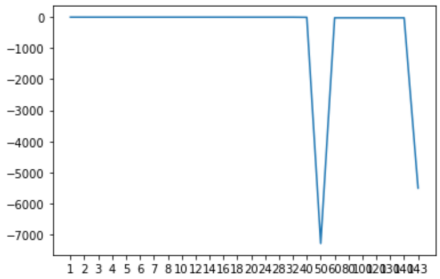
3.3 Lasso Regression

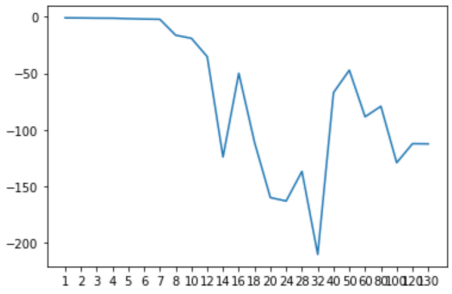
- Initial Cross Validation results with Normalization to regularize dataset

- using SelectKBest:


0.0
−0.5
−1.0
−1.5
−2.0
−2.5
1 2 3 4 5 6 7 8 10 12 14 16 18 20 24 28 32 40 50 60 80 100 120 130


1e8
0.0
−0.5
−1.0
−1.5
−2.0
Accuracy
0
200
400
600
800
1000
CV


0
−1000
−2000
−3000
−4000
−5000
−6000
−7000
1 2 3 4 5 6 7 8 10 12 14 16 18 20 24 28 32 40 50 60 80 100 120 130 144 3


0
−50
−100
−150
−200
1 2 3 4 5 6 7 8 10 12 14 16 18 20 24 28 32 40 50 60 80 100 120 130

3.4 Random Forests

- After removing features with low correlation with target variable (—x—¡0.001) and removing features with low variance, we get these scores for selectKBest features:

- After using Grid Search for Hyper Parameter tuning, we get these scores for selectKBest features