Tesla Stock Price Prediction

VE406 Apply Linear Regression using R

Boqun Li, Xinmiao Yu, Shensong Zhao

UM-SJTU Joint Institute

Abstract

In this project, our purpose is to predict the stock price of Tesla. We want to collect data relevant with the stock price from the Internet. Then analysis will be done to find out underlying problems including the correlated error, etc. Different models and methods will be used to address these problems. Also, variable selection will be done when we are fitting these models on the training dataset. In the end, we will compare among the models and evaluate them with a specific score based on the testing dataset. In this way, we will be able to choose the best model.

Introduction

This project aims to analyze the Tesla stock price and collect related data to do regression analysis and predict $12.7 - 12.11$ stock close price. Our overall flow chart is show in 1.

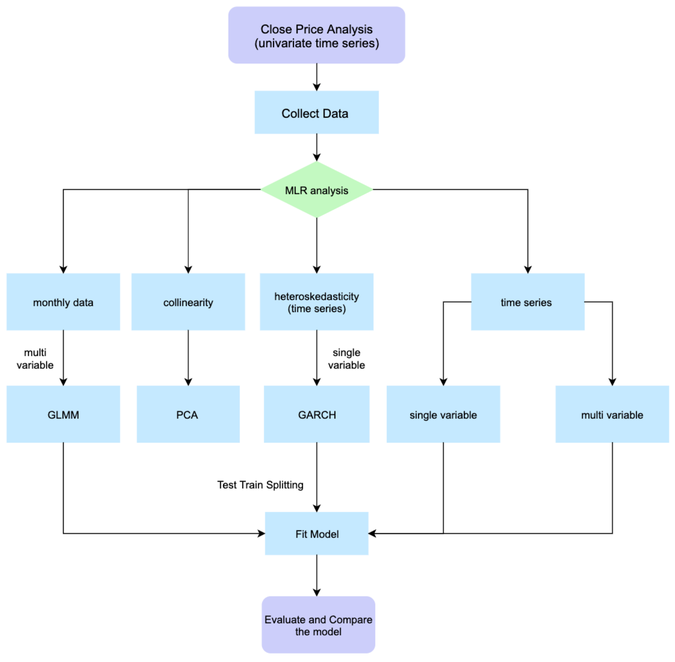

Figure 1: Fit *Close* Flow chart of this project

Close Price Analysis

First analyze as a univariate time series.

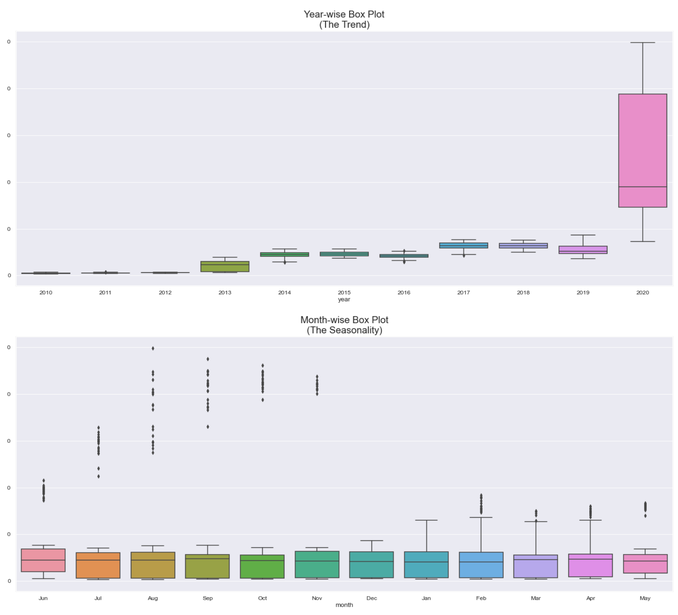

Figure 2: The year-wise and month-wise box plot for potential trend and seasonality. Observe a sharp increase in year 2020 and no obvious seasonality presented. Outliers in the month-wise plot will disappear after fit the same plot with year 2020 data only.

- As goal is predict $12.7 - 12.11$ stock price, decide to **only use year 2020 data**.

Data Collecting

The data or variables we used are shown in the table

oil.csv	death.csv	DPRIME.csv	TOTALSA.csv
OilPrice, daily	Deaths, daily	DPRIME daily	TOTALSA monthly data divided by 30, assgin to each day
tesla-death.csv	**COVID-19-death**	**COVID-19-new-cases**	**GoogleTrend.csv**
the death numbers caused by Tesla	the daily number of death caused by COVID-19	the new cases for COVID-19, daily	The number of people searching for TSLA on Google

Figure 3: Variable or data we used.

MLR Analysis

Variable Selection

- To find the significant regressors for the multi linear regression, we need to provide the R squared, adjust R Squared, Mallow's Cp, and the BIC for different variable.

- From the figure, we need to select the variables with dark colors. So it is easy to see that the deaths, new_deaths, and volume are not significant for the linear regression model.

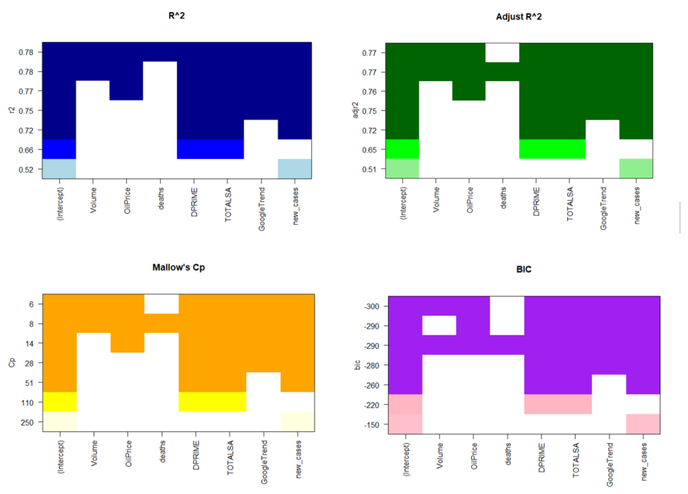

Figure 4: output of regsubsets.out plot for different variables to select the significant variable to the model

- There are some problems in the multi linear model.

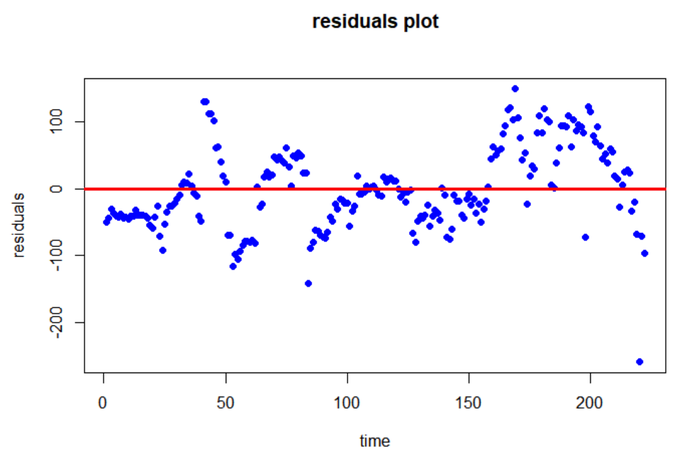

Figure 5: The residual plot. The variance of residuals increases with the date, and there exists Heteroskedasticity.

Moreover, from the acf plot we can see that the errors are highly correlated with each other until the lag around 10, so we also need to analysis the price of the stock as a time series.

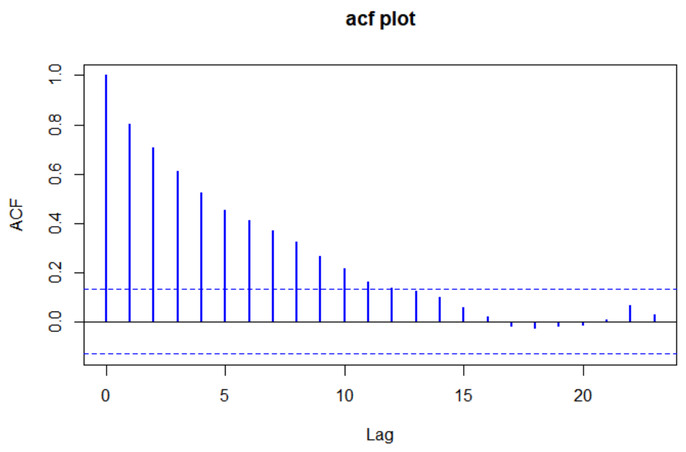

Figure 6: Acf plot shows the errors are highly correlated with each other until the lag around 10, the price of the stock should be analyzed as a time series.

Problem Addressing

- For monthly data and heteroskedasticity, GLMM and GARCH are used respectively. No significant influence shown from these metho ds.

Colinearity

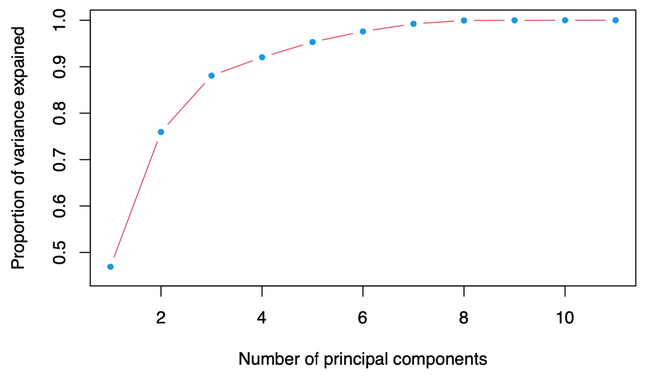

Figure 7: he Cumulative Proportion of Variance Explained by Different Principal Component.

The first four principal components explain over 0.9 of the total variance. Hence we consider the first four components are enough for the model fitting.

Time Series

- Single *Close* price model fit, including *simple Exponential Smoothing (SES),Holt's Method with linear and exponential trend, Holt-Winter expoential trend with addition-addition and addition-multiplication(HWES add-add, HWES add-mul), Seasonal Autoregressive Integrated Moving Average (SARIMA) and autocorrelation(AR).*
- Multiple variable fit. Use *GLMM, ARIMA, Vector-AutoRegression (VAR)* with $4, 5, 3$ varibales respectively.
- All model prediction score are calculated and compared in later section.

Model Score Comparison

- **Method**: Train/Test Spilt
- **Criteria**: Predict test set value is y_{pred}, original test set y_{test}, calculate the score $\sum(y_{pred} - y_{test})^2$.

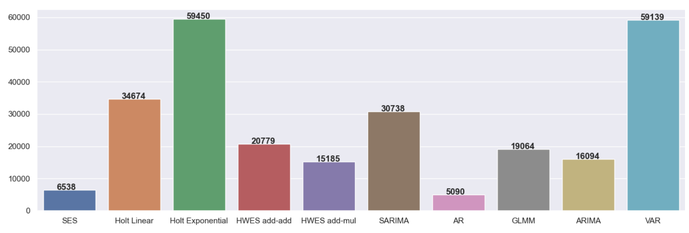

Figure 8: Test score for each model. The smaller the score, the better the model is. So choose *AR* model.

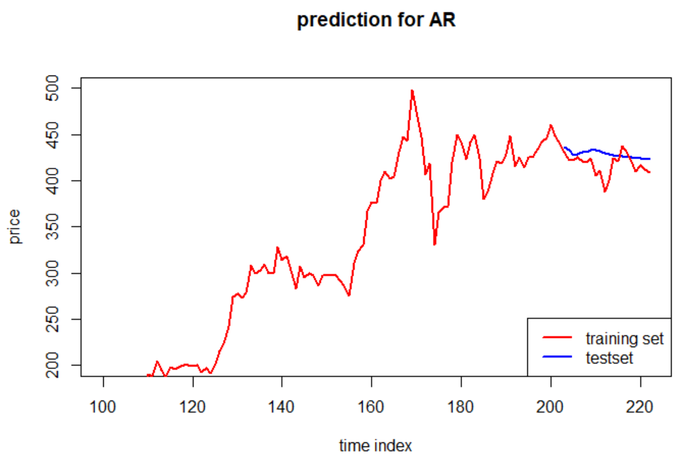

Figure 9: Final AR Model Prenset

Conclusions

We basically find extra parameters that can be used as regressors for stock price prediction. MLR is used to do basic variable selection and problem identifying. We find that there are four problems in total and different methods are used to address them. Then we do predictions of different models on the testing dataset. A bar chart is plotted to compare the residual scores. AR model is finally picked due to its smallest residual sum of squares. Therefore, we use AR model to do the final prediction of stock closing price.