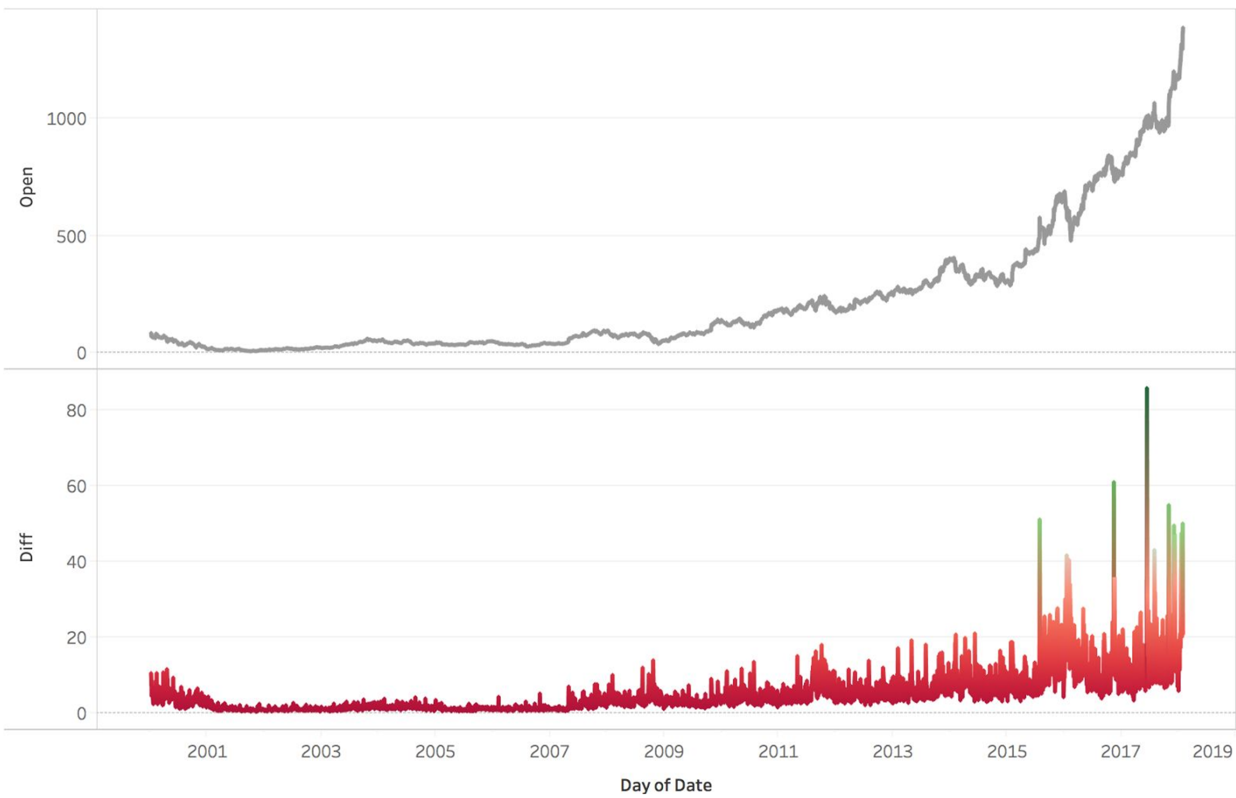
Can we predict the Daily Volatility of Amazon Stock Prices for the next month?

- Maragatham K N

Problem Statement

Amazon has a strong hold in stock market and has experienced a tremendous growth in its stock prices. There has been a whopping increase of 850%.

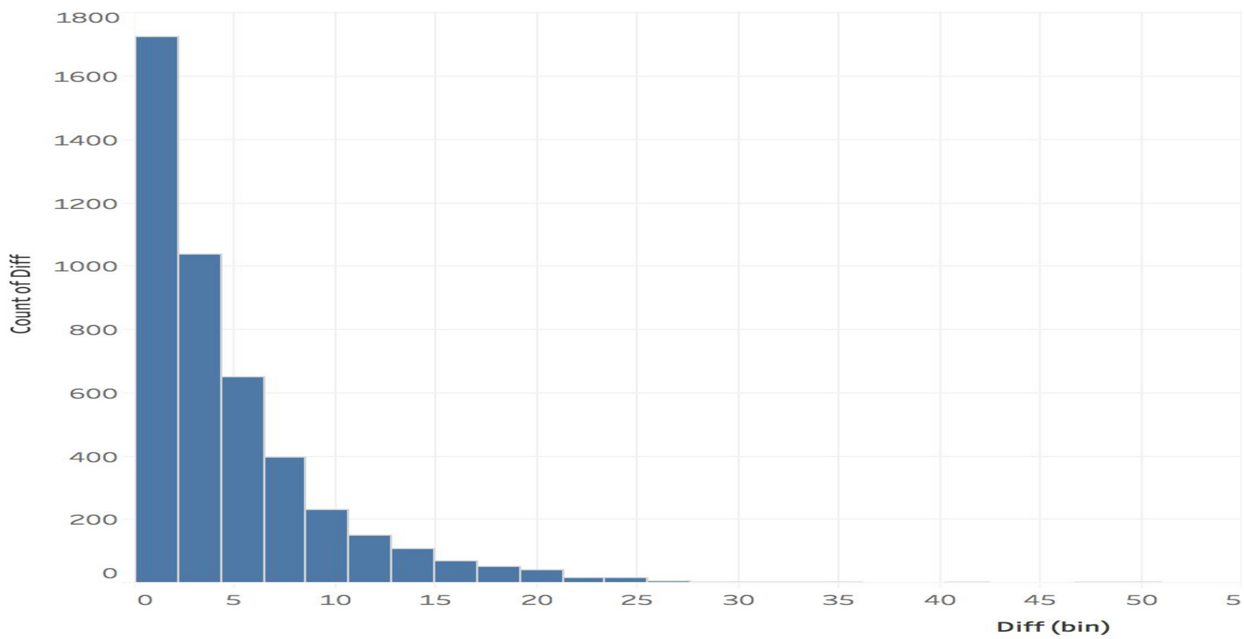


The daily volatility also has been varying across the period.
The problem is to predict the daily volatility change. This could help in giving the companies an insight of how and what factors are affecting this change. This could also help the investors who would know exactly when to invest and make profit or when to sell the stocks to make more profit.

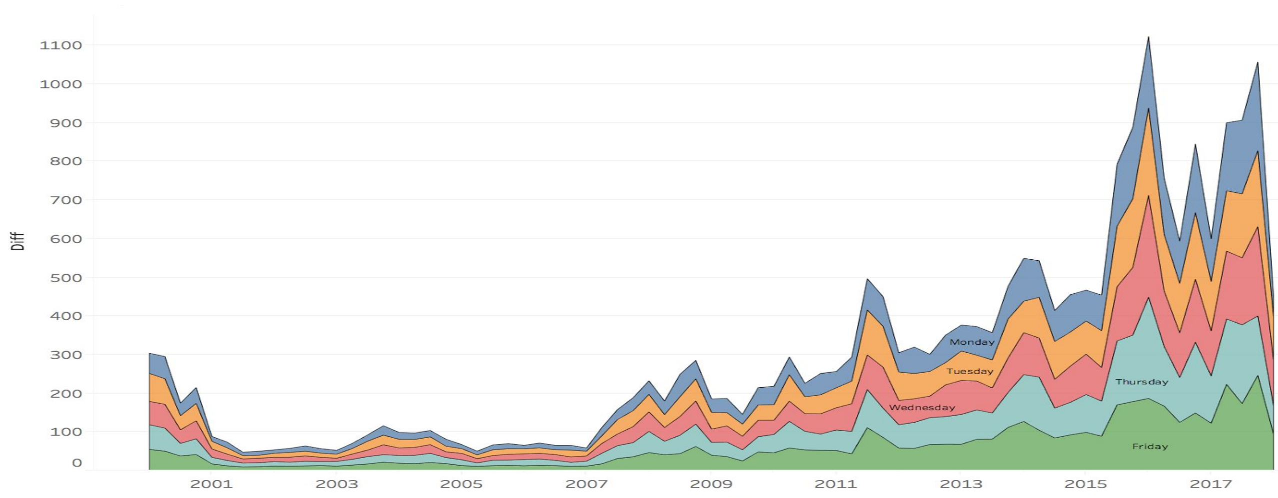
Exploratory Data Analysis:

The daily stock details has been scraped from Yahoo finance using Selenium. The extra fields like quarterly GDP, Revenue, EPS were obtained from NASDAQ and merged with the data.

The exploratory analysis reveals that the Daily Volatility is not normally distributed. It is skewed with data varying between 0-60$.



The stock data is uniformly distributed between the days of the week indicating



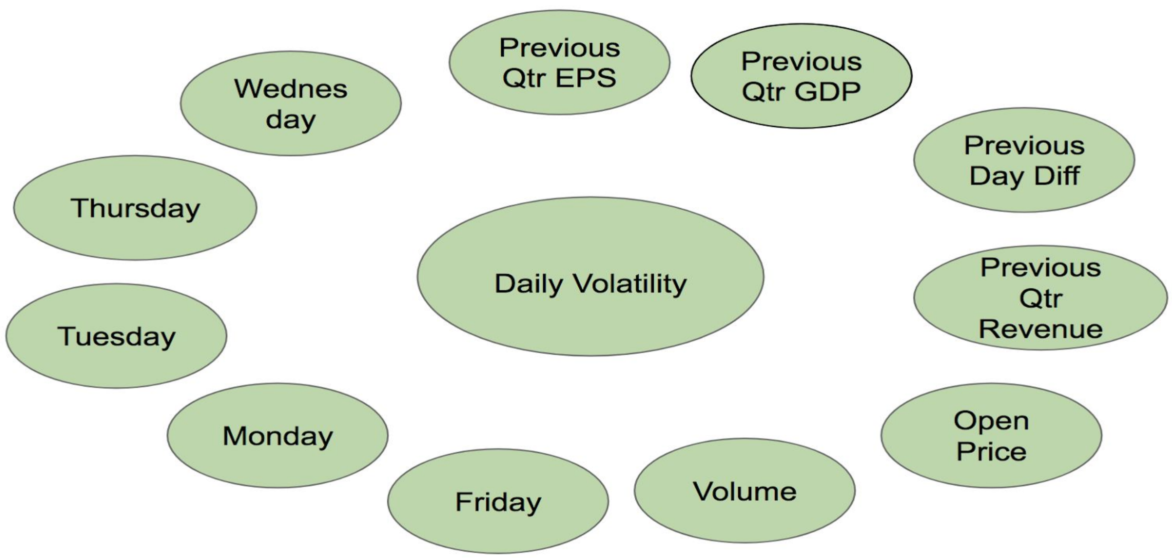
Even when there is a dip in the daily volatility in a week as we can see during the periods 2001-2007, the daily volatility is uniformly divided amongst the days. This indicates that Daily volatility may not be affected by day of the week.

Tools

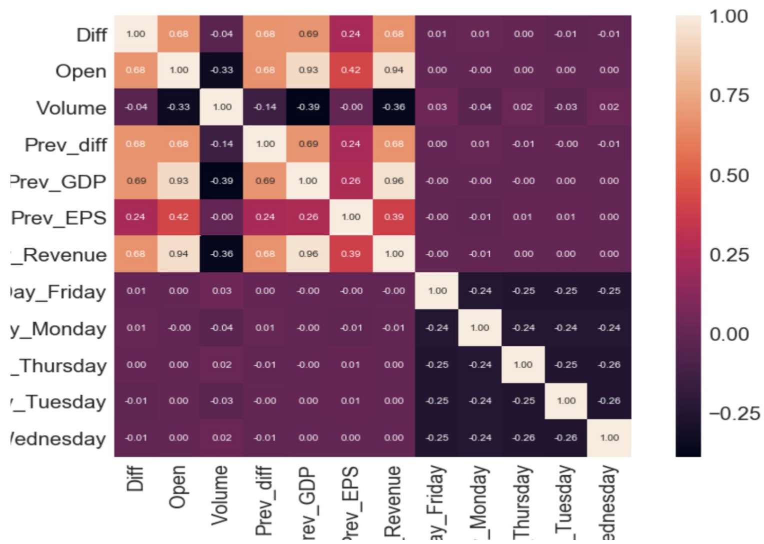
Pandas and Tableau

Approach

Feature Engineering



Correlation plot indicate that target has moderately high correlation with the Previous quarter's GDP, Previous daily volatility and previous quarter's revenue and the Opening stock price



Baselining

Started creating model with the feature Previous quarter’s GDP and kept on increasing features to get an adj-R-squared of .654 with 5 features

- Previous Day’s Daily Volatility
- Open Stock Price
- Volume
- Previous Quarter’s GDP
- Previous Quarter’s Revenue

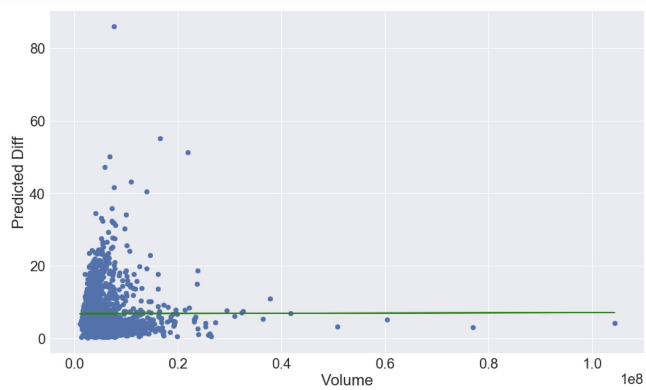
OLS Regression Results

Dep. Variable:	Diff	**R-squared:**	0.655
Model:	OLS	**Adj. R-squared:**	0.654
Method:	Least Squares	**F-statistic:**	782.6
Date:	Wed, 31 Jan 2018	**Prob (F-statistic):**	0.00
Time:	21:06:37	**Log-Likelihood:**	-5289.4
No. Observations:	2070	**AIC:**	1.059e+04
Df Residuals:	2064	**BIC:**	1.062e+04
Df Model:	5		
Covariance Type:	nonrobust		

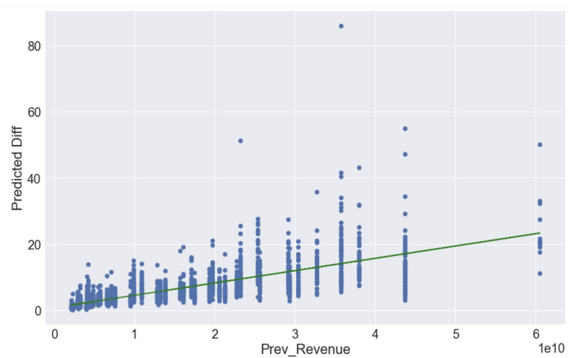
	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	-15.6135	2.281	-6.844	0.000	-20.087	-11.140
Prev_diff	0.3576	0.017	21.523	0.000	0.325	0.390
Prev_GDP	1.084e-12	1.62e-13	6.707	0.000	7.67e-13	1.4e-12
Open	0.0006	0.001	0.753	0.451	-0.001	0.002
Volume	1.711e-07	1.38e-08	12.371	0.000	1.44e-07	1.98e-07
Prev_Revenue	6.631e-11	2.43e-11	2.733	0.006	1.87e-11	1.14e-10

Omnibus:	1472.764	**Durbin-Watson:**	2.053
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	74193.310
Skew:	2.777	**Prob(JB):**	0.00
Kurtosis:	31.799	**Cond. No.**	5.38e+14

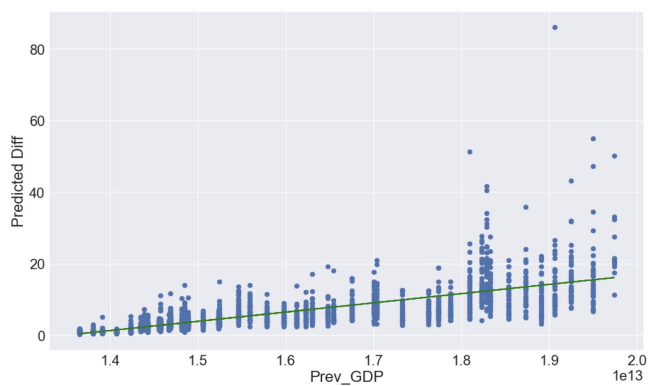
I analyzed the residual plots, the regression fit plots and the Q-Q plot in cross validation mode for linear regression and polynomial regression (degrees 2-9)



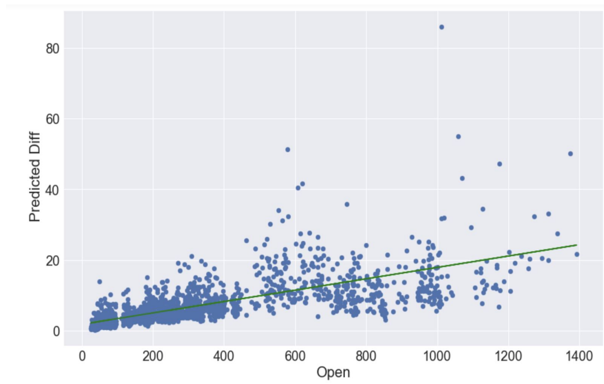
Predicted Diff
80
60
40
20
0
0.0
0.2
0.4
0.6
0.8
1.0
Volume
1e8

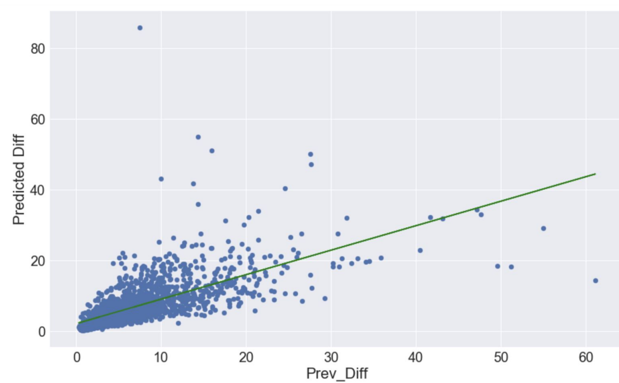


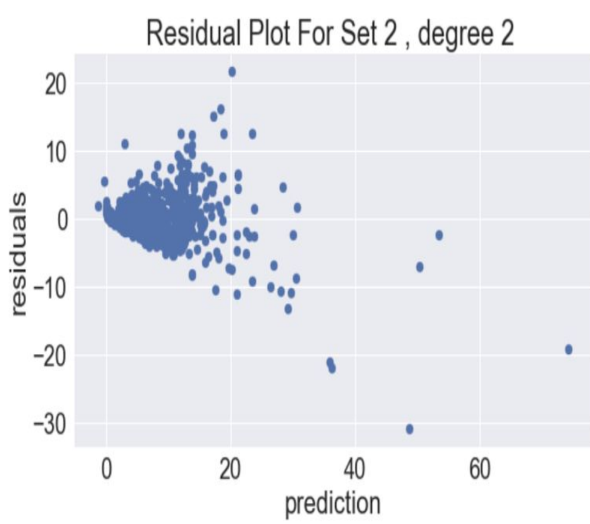
Predicted Diff
80
60
40
20
0
0
1
2
3
4
5
6
Prev_Revenue
1e10

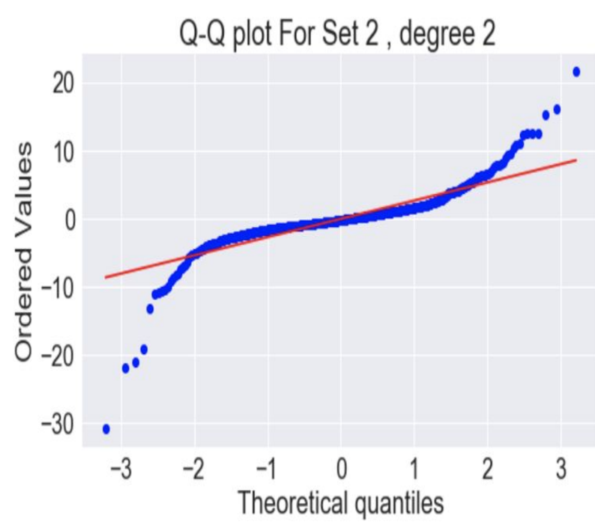


Predicted Diff
80
60
40
20
0
1.4
1.5
1.6
1.7
1.8
1.9
2.0
Prev_GDP
1e13









After analysis I selected degree 2 polynomial regression based on the score.

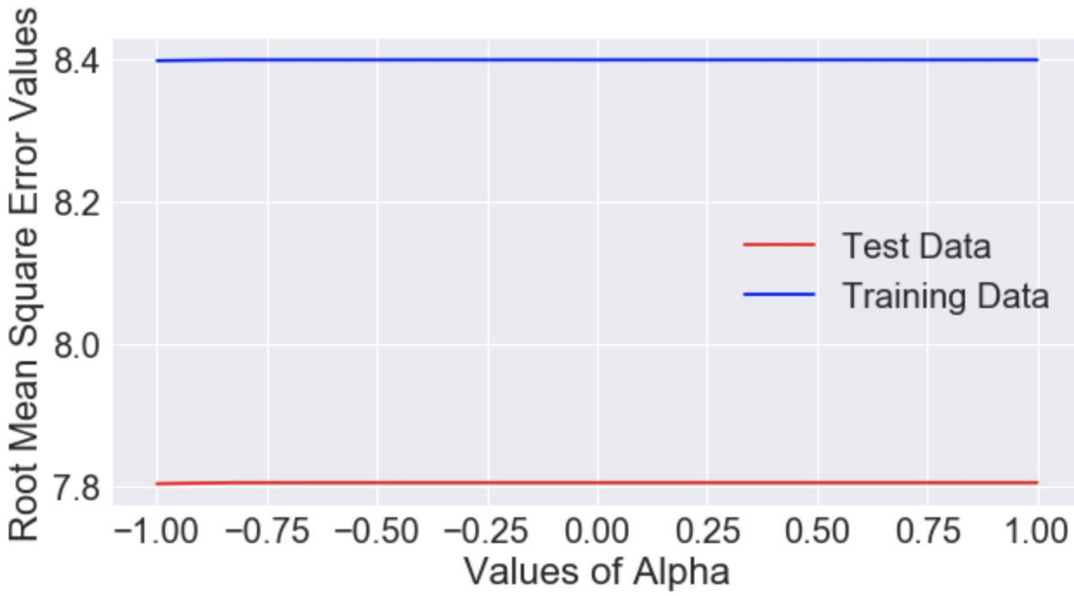
Preliminary Results

The test set fared a negative R^2 and a huge RMSE indicating that something is wrong with the model.

I checked the Random Forest regressor and the Gradient Booster regressor and found that the features Previous Quarter's revenue and Open price are not that important so I finalized a polynomial regression model of degree 2.

After testing, the difference in RMSE for Test and Training data was .7

```
Mean of Error for Test Data :7.805509267545955
Mean of Error for Training Data:8.398808077540318
Alpha: 0.1
Alpha : 0.1
```



Still the Mean Absolute Percentage error was around 78% indicating that this model is not without flaws and cannot be used to predict the Daily Volatility but it does give an insight so as to how the features interact and affect the Daily Volatility. The day of the week particularly has no hand in the Daily Volatility.

What else we could do

Caveats

The number of features deciding the Daily Volatility could be anything ranging from the PE ratio to the company earnings, that could be incorporated.
To predict we need to take the inflation into consideration to mitigate the skew or we could take really short time periods like a month or a week and take all the ticker data(the changes in stock prices per sec) to understand the trend.
Any announcement related to the company could affect the change.

Conclusion

The model is still not in the perfect shape to predict the Daily Change Volatility. It just gives us the insight of the features used.

Appendix

Data Cleaning

I dropped the outliers to mitigate the skew but it still prevailed in the residual plot.