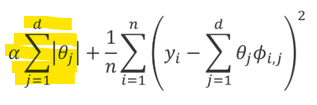
**Lasso Regression features selection with K-fold Cross-Validation: A Statistical model to predict 2-days net change in stocks prices.**

**Background:**

Least Absolute Shrinkage and Selection Operator (Lasso), is a type of linear regression that adds a penalty to the standard sum of squared residuals used in ordinary least squares (OLS) regression.



This penalty, often referred to as L1 regularization, helps to address some common issues in linear regression models including handling Multicollinearity by shrinking the coefficients of correlated variables towards zero. Also, Lasso prevents Overfitting, which occurs when a model learns too much noise from the training data, by penalizing the absolute size of the coefficients (L1 penalty) in high dimensional datasets.

Furthermore, Lasso offers several advantages including but not limited to: Features Selection in high-dimensional data and improving model interpretability as it focuses only on the most relevant predictors from a large pool of predictors that have a meaningful impact on the target variable.

In machine learning and statistics, K-fold cross-validation is a resampling technique used to evaluate the performance of a model. It involves splitting the dataset into K subsets (or folds) of approximately equal size. The model is trained K times using different subsets of the data as the training set each time. For each iteration, one of the K subsets is used as the validation set, and the remaining K-1 subsets used as the training set, each time computing the mean squared error, a performance evaluation metrics.

K-fold cross-validation helps in understanding the bias-variance trade-off. By observing the model performance across different folds, one can assess whether the model is underfitting (high bias) or overfitting (high variance) the data. Also, in research settings, K-fold cross-validation is preferred because it provides a more accurate, reliable and comprehensive evaluation of the model's performance and ensures that the findings are more likely to generalize to new data.

In this study I used Lasso Linear Regression with K-fold Cross-Validation as a features selector to identify most relevant predictors used to fit a linear regression predictive model on stocks price changes data. Two-days stocks price change was chosen because of the high volatility in the stock market making long term predictions hard.

**Data preparation**

**Source**: Data downloaded from barchart.com and processed was used. N = 44610

**Variables:**

Index(['y\_netdiff2', 'net\_diff1', 'symbol', 'extractdte', 'yyyymm', 'mth',

'wk', 'dowk', 'domth', 'last', 'key\_chartp', 'rsi\_3d\_n', 'rsi\_3d\_pchg1',

'rsi\_9d\_n', 'rsi\_9d\_pchg1', 'rsi\_9d\_pchg2', 'rsi\_9d\_pchg5', 'rsi\_20d\_n',

'Pcnt\_stoch3', 'Pcnt\_k\_stoch\_d', 'Pcnt\_di14d\_up\_above\_dn',

'Pcnt\_5d\_mvavg\_50d', 'Pcnt\_Above\_mvavg\_5d', 'Pcnt\_Above\_mvavg\_50d',

'Pcnt\_Above\_mvavg\_20d', 'Pcnt\_Above\_mvavg\_200d', 'Pcnt\_Above\_bb20dl',

'Pcnt\_Above\_AvgResist', 'Pcnt\_Above\_AvgPx\_stock', 'crit9\_52wk\_ratio',

'pt\_below\_52wk\_high', 'pt\_above\_floor', 'pt\_below\_ceil',

'pt\_hammer\_above\_floor', 'pchg\_floor\_min1', 'pchg\_floor\_min0',

'pchg\_ceil\_max0', 'macd\_9d\_n', 'macd\_14d\_pchg2', 'macd\_14d\_n',

'hvol\_cat', 'Fib\_382pcnt\_n', 'dcr', 'days\_since\_volspike',

'days\_since\_target\_raised', 'days\_since\_hvolb\_1mth',

'days\_since\_highprob\_bull', 'days\_since\_drop\_8',

'days\_since\_bull\_window', 'days\_since\_bull\_strength',

'days\_since\_bull\_stoch', 'days\_since\_bull\_rev',

'days\_since\_bull\_hammer', 'days\_since\_bear\_countered', 'days\_since\_bbl',

'bband\_width', 'atr\_3d\_n', 'atr\_9d\_n', 'atr\_14d\_n'],

dtype='object')

**Cleaning**: No duplicated records. Variables with missingness percentage >=20%

['pt\_hammer\_above\_floor','hvol\_cat','days\_since\_hvolb\_1mth','days\_since\_bear\_countered'] were dropped from the dataset while the rest were imputed using the variable’s median value.

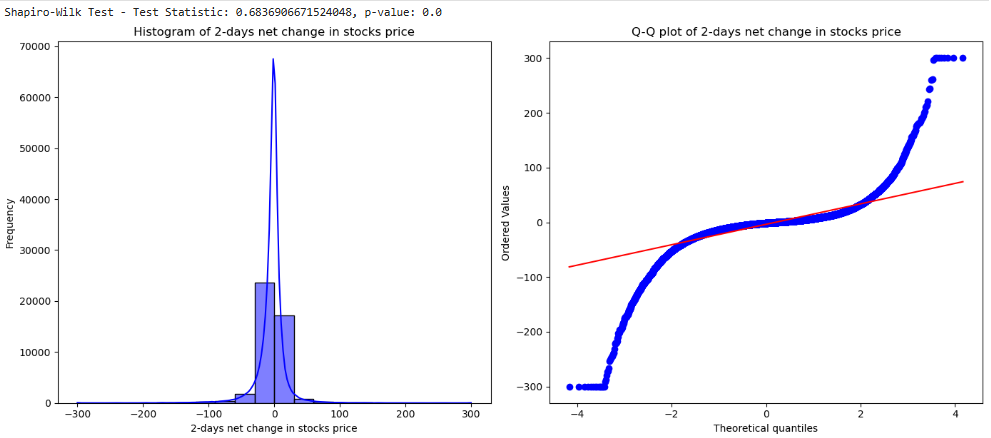
**Transformations:** Polynomial features to the second degree (squared) where created on this multidimensional dataset resulting into 1377 columns.

**Data splitting**: Dataset was split into a training (80%) and test (20%) set.

print(X\_train.shape, X\_test.shape, y\_train.shape, y\_test.shape) >> (35688, 1377) (8922, 1377) (35688,) (8922,).

**Data Modelling**

**Normality:** Q-Q plot and Shapiro-wilk Test revealed non-normality of 2-days net change in stocks price data (y\_netdiff2).



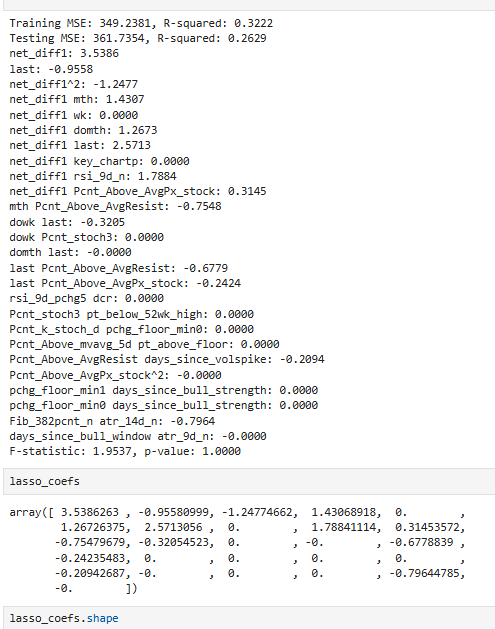
**Shapiro-Wilk test**: H0: The data are normally distributed. H1: The data is not normally distributed.

Since the p-value < 0.05, we reject the null hypothesis and conclude that there is sufficient evidence that the data is not from a normally distributed population.

**Outliers:** From the scatter plot (in appendix) a few outlying records that need investigation are observed. For this study, I did not drop them.

**Multicollinearity:** Inflated and unstable coefficients due to high collinearity in high dimensional datasets are penalized by Lasso. Lasso effectively perform feature selection by shrinking the coefficients of less important variables to zero. This helps in identifying the most relevant predictors for the target variable. Using GridSearchCV with K-fold Cross-Validation (k=5), the best tuning/regularization (alpha) value selected to control the strength of the penalty applied to the coefficients was 1.87.

**Dimension reduction/Feature importance:** Lasso automatically reduced the predictor space from >1,000 predictors to 26 predictors with 14 non-zero/important predictors below:



**Final model:**

Dropping non-significant higher dimensional/interaction terms [*'net\_diff1\_sq', 'last\_avgresist'*] I re-fitted the final model below:

R-squared: 0.36520753399844397

Adjusted R-squared: 0.36463780888778863

F-value: 641.0241135032397

F-test p-value: 0.0

**Equation**: y = -2.6583 + (0.3621 \* net\_diff1) + (-0.0043 \* last) + (0.0007 \* Fib\_382pcnt\_n) + (-0.1542 \* mth) + (-0.0229 \* domth) + (-0.1228 \* dowk) + (0.2640 \* Pcnt\_Above\_AvgResist) + (-0.0075 \* atr\_14d\_n) + (0.0108 \* Pcnt\_Above\_AvgPx\_stock) + (-0.0393 \* days\_since\_volspike) + (0.0428 \* rsi\_9d\_n) + (0.0098 \* Pcnt\_5d\_mvavg\_50d) + (-0.0037 \* net\_diff1\_sq) + (0.0149 \* net\_diff1\_mth) + (0.0062 \* net\_diff1\_domth) + (0.0002 \* net\_diff1\_last) + (0.0033 \* net\_diff1\_rsi) + (0.0042 \* net\_diff1\_avgpx) + (-0.0875 \* mth\_avgresist) + (-0.0003 \* dowk\_last) + (-0.0008 \* last\_avgresist) + (-0.0001 \* last\_avgpx) + (-0.0881 \* avgresist\_ds\_volspike) + (-0.0001 \* Fib\_atr)

Mean Squared Error (MSE) on Validation Set: 358.5732

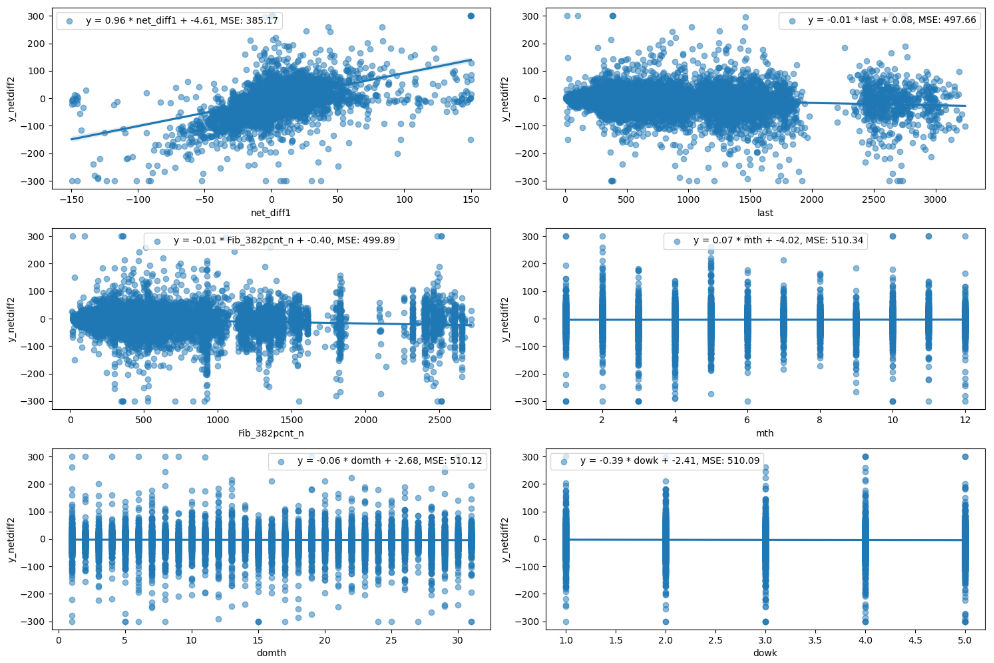
Mean Squared Error (MSE) on Test Set: 363.3694

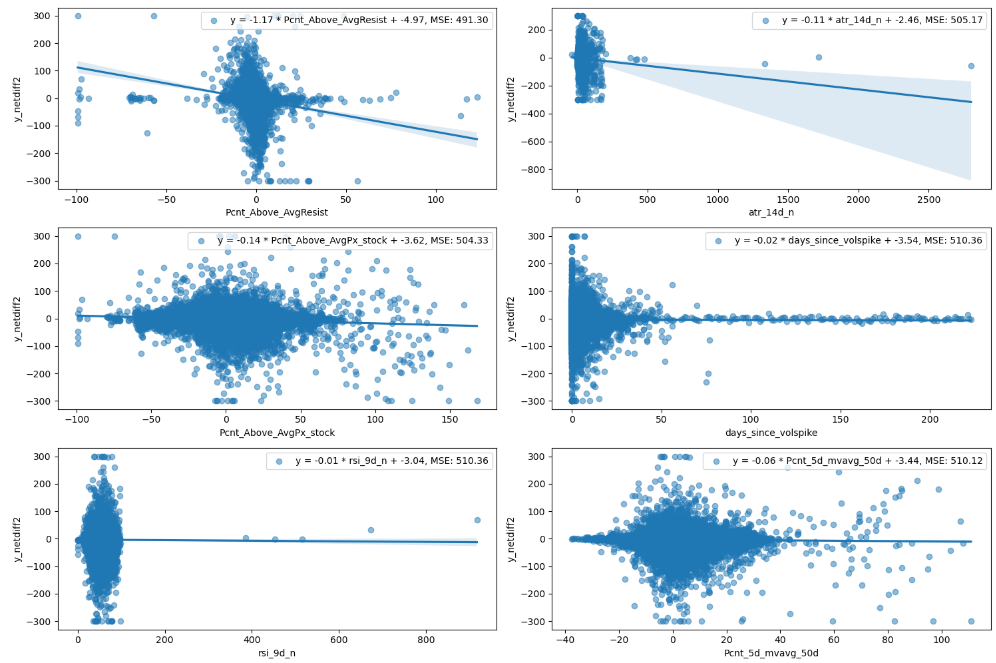
**Interpretation:**

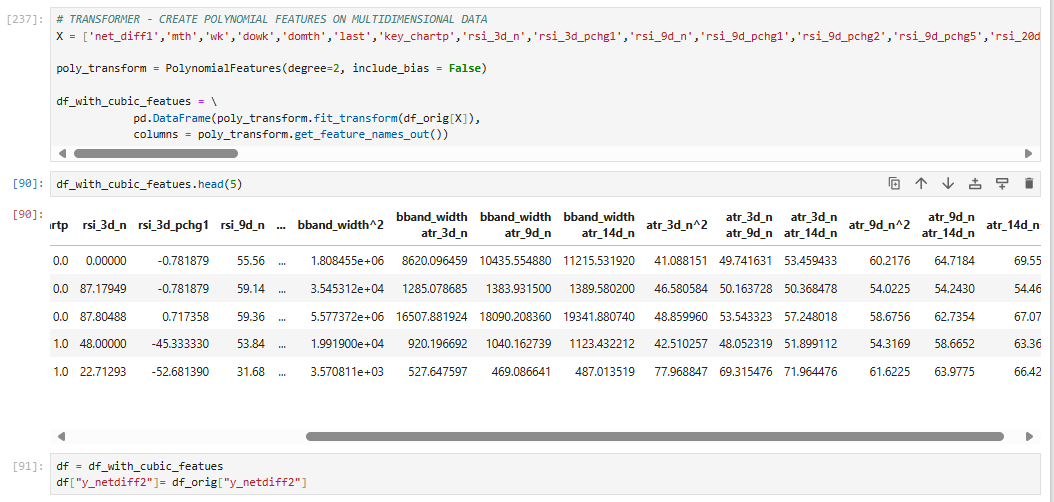
The overall model is significant (p-value for the F test <0.0001)

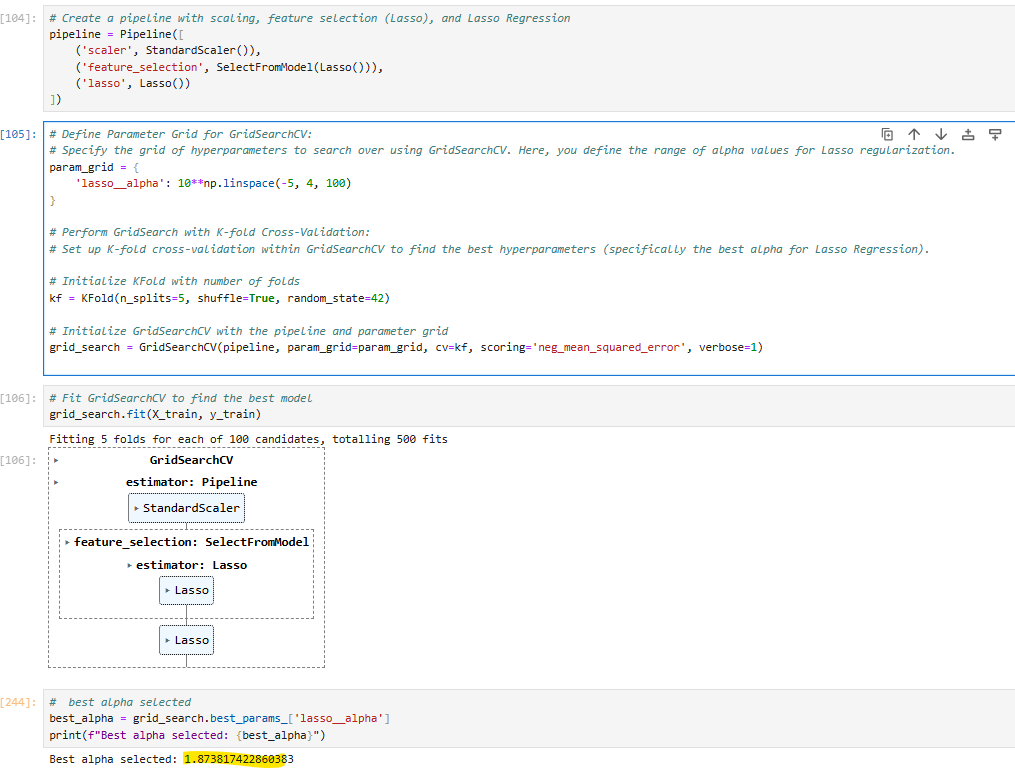
Adjusted R-squared= 0.37 meaning that only 37% of the variance in the 2-days net change in stocks prices is explained by the variables in the model.

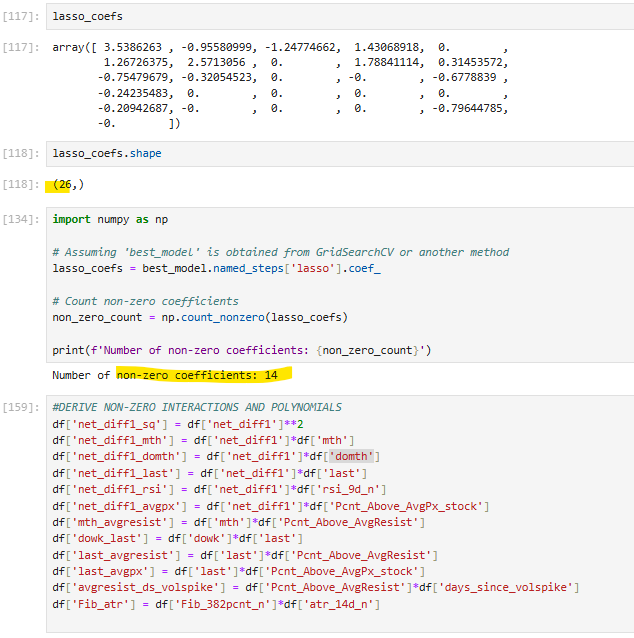
**Appendix:**

****

****

****



****

**Limitations:**

More computing power is required for higher dimensions especially for stocks data.

Not enough time to explore interactions and their interpretations as well as other model diagnostics.