Project Final Report

Project Title:

"Money Left of the Table:" A Macroeconomic Approach to Predict IPOs Initial Return

Problem Statement:

As of June 15, 2021, with more than six months to year end, U.S. Initial Public Offerings (IPOs) totaled $171 billion and have already broken the 2020 record of $168 billion [1], in part, due to the increasing trend of Special Purpose Acquisition Company (SPACs) [9]. Such a high volume of IPOs raises the question of how much money was "left on the table?" IPO initial return or "money left on the table" refers to the IPO underpricing and has been puzzling both researchers and practitioners for decades [10]. Ritter [12], back in 1987 reported that, on average, investors who purchased IPO stocks at the offer price have earned 14.8% at the end of the first day of trading. In 2020, average initial return was 34% for non-tech and 63.7% for tech IPOs [14]. Moreover, in a paper documenting IPOs underpricing across 54 countries around the globe, Loughran, Ritter and Rydqvist [6] found an average country-level of underpricing ranging from 3.3% (Russia within 1999-2013) to 270.1% (United Arab Emirates within 2003-2010).

Initial return which is the difference between the offering price and day-one closing price is measured in this paper as

$$R_i = (P_{ic} - P_{i0})/P_{i0}$$

where P_{i0} is the offering price of stock *i* and P_{ic} is the closing price of stock *i* on the first trading day, i.e. the day the company was taken public. The underpricing phenomenon begs the question of why issuing firms sell their shares, which is equivalent to ownership or interest in the firm, for a discount and leave "money on the table?" Many theories, such as information asymmetry, institutional explanations, ownership and control reasons, and behavioral explanations, among others, have been suggested in attempt to rationalize and explain the illogical practice [5] and as indicated by Quintana et al., is still a very active field of investigation [10].

In attempt to predict IPOs initial return, Ordinary Least Squares (OLS) Linear Regression has been applied for decades and still is the dominant approach in predicting initial returns [2]. Luque et al. [7] and Huang et al. [4] focused on predicting IPO underpricing using Genetic Algorithms. Baba and Sevil [2], and Quintana et al. [10] predicted underpricing using Random Forest. Mitsdorffer and Diederich [8] predicted with Artificial Neural Networks (ANN) and Support

Vector Machine (SVM). Robertson et al. [13], Wang et al. [15] and Reber et al. [11] also attempted to predict initial returns using ANN.

While various machine learning algorithms have been applied to predict IPOs initial return, all researches have used only "firm related" or offering structure [10] predictors such as Net Income, Earnings per Share (EPS), firm size, offer size and various financial ratios. Since trying to predict initial returns using the same methodologies would be nothing but replicating the work already done, in this paper we will attempt to improve the prediction accuracy by including macroeconomics indicators as predictors together with "firm related" ones. Since linear regression has been the most common approach in predicting IPOs initial return, we will compare prediction accuracy and Adjusted R^2 between linear regression with only "firm related" predictors (the base model) with linear regression with the same "firm related" predictors and macroeconomics indicators (the complete model). In previous work, many different "firm related" and offering structure type of predictors have been used and including all of them would be impractical and would probably cause to overfitting. Hence, the approach taken is that assuming the complete model performs better than the base model, the conclusion would be that previous work could have improved their models' accuracy by introducing macroeconomic predictors in addition to the ones used, and that future work should consider to include such "exogenous " type predictors.

The rationale behind introducing macroeconomic indicators as potential predictors is based on the hypothesis that the information hidden in macroeconomic variables alone can be used to accurately predict stocks prices and sector indices. Weng et al. [16] compared the prediction accuracy of major U.S stock and sector indices between different machine learning algorithms using only macroeconomic variables with prediction using time series algorithms such as Auto Regressive Integrated Moving Average (ARIMA) and Generalized AutoRegressive Conditional Heteroskedasticity (GARCH) and states that "the use of macroeconomic indicators (alone, via an ensemble) are more predictive than the information contained in historical prices (alone)" (as used in the time series approach where past observations are used to predict future observations).

The approach suggested in this paper is novel because predicting IPOs initial return using macroeconomic variables, or investigating the predictive power of macroeconomic indicators, to the best of our knowledge, has never been done in the past. In addition, while Weng et al. concludes that macroeconomic indicators can predict stock prices better than other "traditional" methodologies, many argue that "industrial production, risk premium change, yield curve twist, and inflation all have significant effects on the variability of stock returns," (i.e. stocks volatility/ risk) "but macroeconomic factors do not have a significant influence on stock price change" [16]. Hence, the paper will investigate whether macroeconomic information can add predictive power in IPOs initial return prediction.

Data Source

The work is based on a sample of 291 companies taken public between May 2014 and May 2021 in the two U.S. stock markets NASDAQ, and NYSE (the New York Stock Exchange). We started with a sample of 2,950 companies taken public between June 2021 and October 2010. We then excluded mutual funds, Pink Sheets, as well as transaction with total value of less than $50 million, and transactions with missing data.

Data about the companies were retrieved from S&P Capital IQ which combines deep and broad global financial intelligence with an array of tools for analysis, ideation and efficiency, and is used by accountants, finance professionals, mergers & acquisition consultants, investment bankers, equity researchers and more. The data include:

- Ticker symbol
- IPO date
- Company name
- Exchange
- Offer price per share before underwriting commission
- First day close price
- Initial return
- "Firm related" variables:
    - Last 12 months revenue at IPO ($mm)
    - Last 12 months EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization) at IPO ($mm)
    - Last 12 months net income at IPO ($mm)
    - Net debt at IPO
    - Last 12 months earnings from continuing operation at IPO ($mm)
    - Last 12 months total cash and short-term investments at IPO ($mm)
    - Last 12 months total assets at IPO ($mm)
    - Technology sector (Categorical/ Boolean variable) – Note that as indicated above, in 2020, average initial return was 34% for non-tech and 63.7% for tech IPOs. Hence, the expectation is that this variable will have predictive power. Companies' sectors were retrieved using the Python 'yfinance' library.

Macroeconomics indicators data were retrieved from the Quandl R package which fetches data from the Federal Reserve Economic Data (FRED), a project by the Economic Research department of the Federal Reserve Bank of St Louis., and include:

- M1 – *weakly data*
- M2 – *weakly data*
- Gross Domestic Product (GDP) – *quarterly data*
- Consumer Price Index for All Urban Consumers: All Items (CPI) – *monthly data*
- Industrial Production Index (IPI) - *monthly data*
- Unemployment Rate (UNRATE) - *monthly data*
- Effective Federal Funds Rate (DFF) – *daily data*

According to FRED (for more information please visit the FRED),

- M1 – Consists of currency outside the U.S. Treasury, Federal Reserve Banks, and the vaults of depository institutions; demand deposits at commercial banks (excluding those amounts held by depository institutions, the U.S. government, and foreign banks and official institutions) less cash items in the process of collection and Federal Reserve float; and other checkable deposits.
- M2 - M1 plus savings deposits; small-denomination time deposits (time deposits in amounts of less than $100,000) less individual retirement account (IRA) and Keogh

balances at depository institutions; balances in retail money market funds (MMFs) less IRA and Keogh balances at MMFs.
- GDP - The featured measure of U.S. output, is the market value of the goods and services produced by labor and property located in the United States.
- CPI – A measure of the average monthly change in the price for goods and services paid by urban consumers between any two time periods. It can also represent the buying habits of urban consumers.
- IPI - Measures real output for all facilities located in the United States manufacturing, mining, and electric, and gas utilities (excluding those in U.S. territories).
- UNRATE - The number of unemployed as a percentage of the labor force. Labor force data are restricted to people 16 years of age and older, who currently reside in 1 of the 50 states or the District of Columbia, who do not reside in institutions (e.g., penal and mental facilities, homes for the aged), and who are not on active duty in the Armed Forces.
- DFF - The interest rate at which depository institutions trade federal funds (balances held at Federal Reserve Banks) with each other overnight.

Methodology

Our starting point is first to fit the base model using only the "firm related" variables. Determining whether the base model is a "good enough" model to be used as a "starting point," we will benchmark it against Hanley's model [3] by comparing our base model's Adjusted R^2 to Hanley's Adjusted R^2 of 17.80%.

We will analyze the base model for multicollinearity by evaluating the Variance Inflation Factor (VIF) with a multicollinearity threshold of *max* (10, $\frac{1}{1-R^2}$) where the R^2 is the coefficient of determination of the regression model including all observations. We will also test whether the regression assumptions hold, which are zero mean, constant variance and independent. More specifically, let ε_i be the error terms, the assumptions are:
- Linearity/ Zero mean: $E(\varepsilon_i) = 0$
- Constant variance: $Var(\varepsilon_i) = \sigma^2$
- Independence: $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are independent random variables
- For statistical inference, we also assume $\varepsilon_i \sim Normal$

In case the normality or the constant variance assumptions do not hold, we would try Box-Cox transformation to improve the goodness of fit.

In attempt to construct a better model than the base one, fitting the complete model (i.e. including both the predicting variables from the base model and the macroeconomics variables) we will go through the same VIF and goodness of fit analyses, as well as variables selection using Ridge regression, Least Absolute Shrinkage and Selection Operator (LASSO) regression and Elastic Net where:
- LASSO: $argmin \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|$ (L1 penalty, "force" some coefficients β_j to equal 0)
- Ridge: $argmin \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}\beta_j^2$ (L2 penalty, shrinks but does not "force" any coefficients β_j to equal 0)
- Elastic Net: $argmin \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j^2$ (enforces some of the regression coefficients β_j to be 0 and shrinks the rest)

Evaluation and Final Results

The entire data was used to both fit the models and prediction. Prediction accuracy will be compared between the base model and the complete model. Measures of accuracy will be performed by:

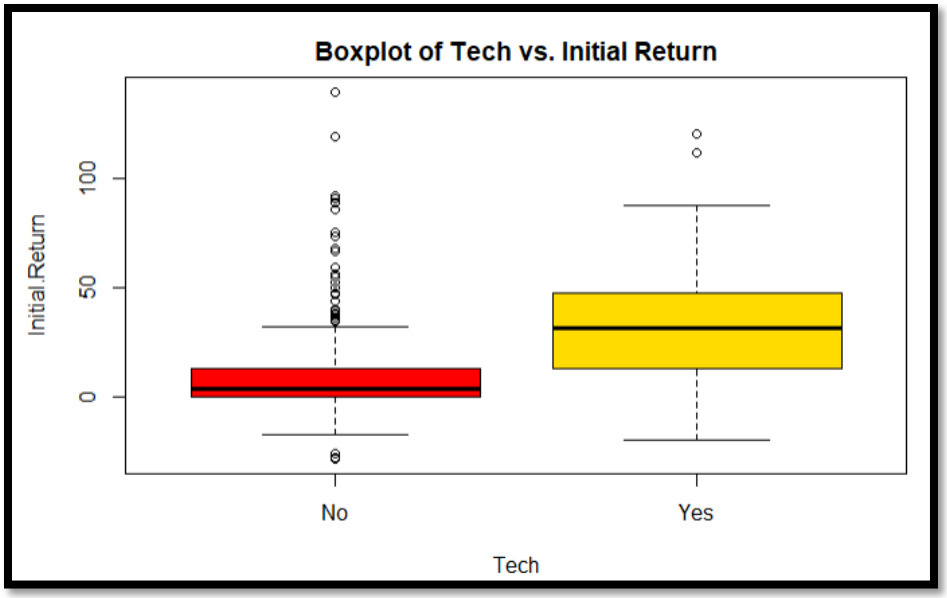
- Mean squared prediction error (MSPE), computed as the mean of the square differences between predicted and observed.
- Precision error (PM), computed as the ratio between the sum of squared residuals and the sum of square differences between the response and the mean of the responses.

MSPE and PM are used because they are appropriate for evaluating prediction accuracy for a linear model estimated using least squares.

The models also will be evaluated by comparing their Adjusted R^2. The coefficient of determination, or R^2, is the proportion of total variability in *Y* (the response variable) that can be explained by the linear regression model, or the amount of variance accounted for in the relationship between two (or more) variables i.e. our response variable Initial Return and the predictors. Simply put, as R^2 increases, the *Y* values would be closer to the regression line (in Simple Linear Regression/ plane in Multiple Linear Regression). R^2 is calculated as:
R^2 = *SSR* / *SST* = 1 - (*SSE* - *SST*), where:
SST = sum of squares total = total deviation
SSE = sum of squared errors = unexplained deviation
SSR = sum of squares for regression = explained deviation
However, if we want to compare models with different number of predicting variables as in our analysis, we should evaluate the Adjusted R^2. Since as we add more predicting variables, the macroeconomic indicators in our case, R^2 can never decrease but only increase or stay the same, we should use the Adjusted R^2, because it is adjusted for the number of predicting variables as it penalizes for the addition of predictors.
Adjusted R^2 is calculated as:

$R^2_{Adj} = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$, where
n is the number of observations
p in the number of predicting variables

As indicated above, in 2020, average initial return was 34% for non-tech and 63.7% for tech IPOs [14]. Hence, since it is expected that industry type (i.e. Tech or non-Tech) would be a significant variable and encompasses predictive power, we started by analyzing whether the initial return mean value is dependent on industry. Consistent with our expectations, since the boxplot below indicates that the mean is different for whether a company is within the Tech industry or not, we expect that the variable will be significant (at least at the 0.05 level) in the regressions model we will perform. More specifically, firms within the Tech industry have higher initial return as they leave more money on the table.



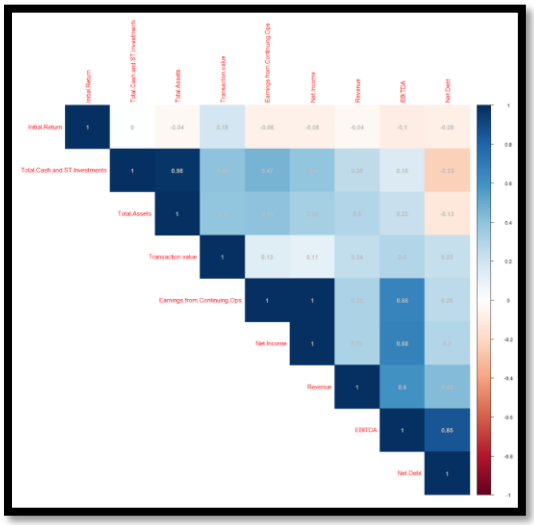
Fitting the base model with all the "firm related" variables, the significant coefficients at the 0.05 level are the intercept, Tech, Transaction value, Total cash & ST investments and Total assets, and has an Adjusted R^2 of 0.1534 (or 15.34%). However, both correlation plot (corrplot) and VIF test indicate a multicollinearity problem. Multicollinearity generally occurs when there are high correlations between two or more predicting variables. In other words, one predicting variable can be used to predict the other. Multicollinearity can cause many problems in the model and its interpretation such as:

1. If one value of one of the *x* predicting variables is changed only slightly, the fitted regression coefficients can change dramatically.
2. It can happen that the overall F statistic is significant, yet each of the individual t statistics is not significant. That is, we will not be able to detect statistical significance because the variance of the estimated coefficients would be artificially large. Another indication of this problem is that the p value for the F test is considerably smaller than those of any of the individual coefficient t tests.
3. Poor prediction of new observations.

Model summary:

```
Call:
lm(formula = Initial.Return ~ ., data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-57.178 -10.556  -6.172   5.195 121.193

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  8.772e+00  1.800e+00   4.872 1.85e-06 ***
Revenue                     -3.391e-06  4.324e-04  -0.008   0.9937
EBITDA                       4.799e-03  1.618e-02   0.297   0.7670
Net.Income                   1.104e-01  1.404e-01   0.786   0.4324
Net.Debt                    -4.576e-04  2.424e-03  -0.189   0.8504
TechYes                      1.958e+01  3.890e+00   5.033 8.64e-07 ***
Transaction.value            7.916e-03  3.840e-03   2.061   0.0402 *
Earnings.from.Continuing.Ops -1.244e-01  1.425e-01  -0.873   0.3833
Total.Cash.and.ST.Investments  2.197e-02  9.023e-03   2.435   0.0155 *
Total.Assets                -1.826e-03  7.393e-04  -2.470   0.0141 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.57 on 281 degrees of freedom
Multiple R-squared:  0.1797,    Adjusted R-squared:  0.1534
F-statistic: 6.838 on 9 and 281 DF,  p-value: 6.797e-09
```

correlation plot:



VIF test:

```
VIF Threshold: 10
                    Revenue                      EBITDA                  Net.Income                  Net.Debt
Tech
                      2.154                      30.219                    1135.673                    18.430
1.046
          Transaction.value  Earnings.from.Continuing.Ops  Total.Cash.and.ST.Investments        Total.Assets
                      1.931                    1223.525                      75.346                    61.189
```

Moving on to fit a base model to address the multicollinearity problem, we selected the variables Revenue, EBITDA, Tech, Net income, Total assets and Transaction value. This model's Adjusted R^2 is 0.1422 and no issues of multicollinearity based on VIF test. The significant coefficients at the 0.05 level are the intercept, EBITDA, Tech, Total assets and Transaction value.

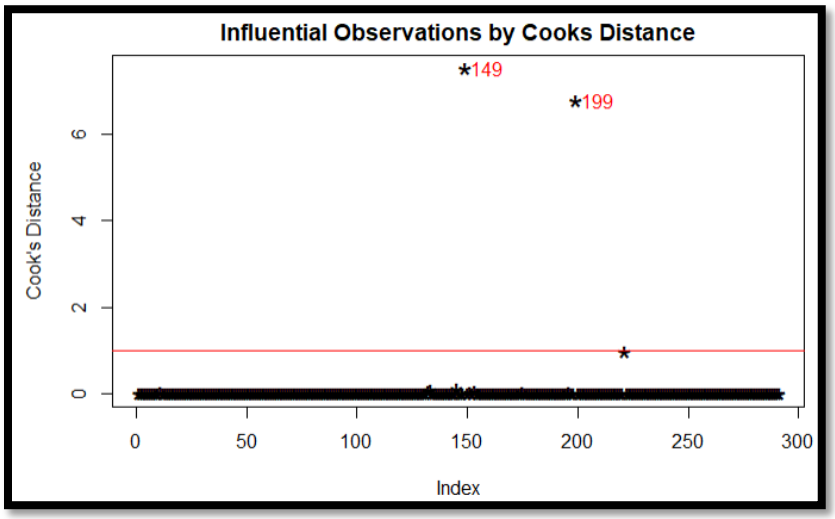
Model summary:

```
Call:
lm(formula = Initial.Return ~ Revenue + EBITDA + Tech + Net.Income +
    Total.Assets + Transaction.value, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-50.463 -10.335  -6.108   3.258 126.988

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.8201357  1.7708566   4.416 1.43e-05 ***
Revenue           0.0003390  0.0003913   0.866   0.3870
EBITDA           -0.0124491  0.0051874  -2.400   0.0170 *
TechYes          20.9454645  3.8723325   5.409 1.35e-07 ***
Net.Income        0.0082527  0.0062885   1.312   0.1905
Total.Assets     -0.0002397  0.0001149  -2.087   0.0378 *
Transaction.value  0.0129069  0.0032488   3.973 9.01e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.72 on 284 degrees of freedom
Multiple R-squared:   0.16,     Adjusted R-squared:  0.1422
F-statistic: 9.014 on 6 and 284 DF,  p-value: 5.033e-09
```

VIF test:

```
VIF Threshold: 10
          Revenue            EBITDA              Tech        Net.Income      Total.Assets Transaction.value
            1.741             3.065             1.023             2.249             1.459             1.364
```

In order to improve the model's fit and prediction accuracy, we removed outliers observations based on Cook's Distance threshold of 1. Two datapoints (indexed at 149 and 199) were flagged as outliers and were removed from the data.



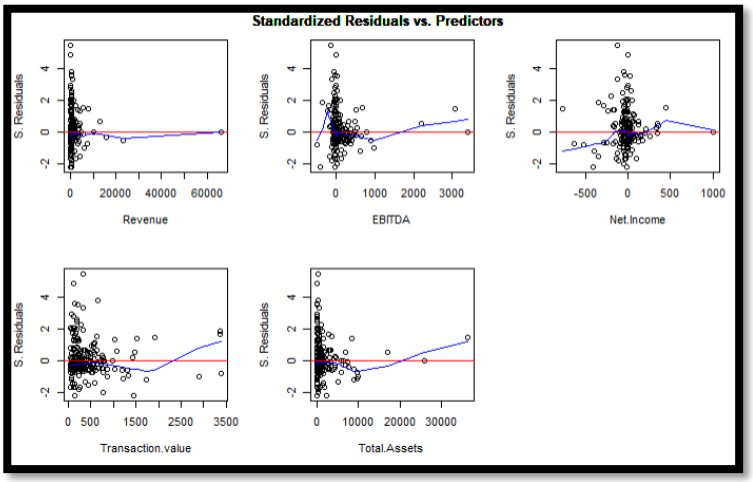
The *final* base model was fitted with the variables Revenue, EBITDA, Tech, Net income, Total assets and Transaction value, after removing the two outlier observations. Now the significant coefficients at the 0.05 level are the intercept, Revenue, Tech and Transaction value with an Adjusted R^2 of 0.1587 which is very close to our benchmark model (i.e. Hanley's model), and an MSPE of 496.88 and a PM of 0.824.

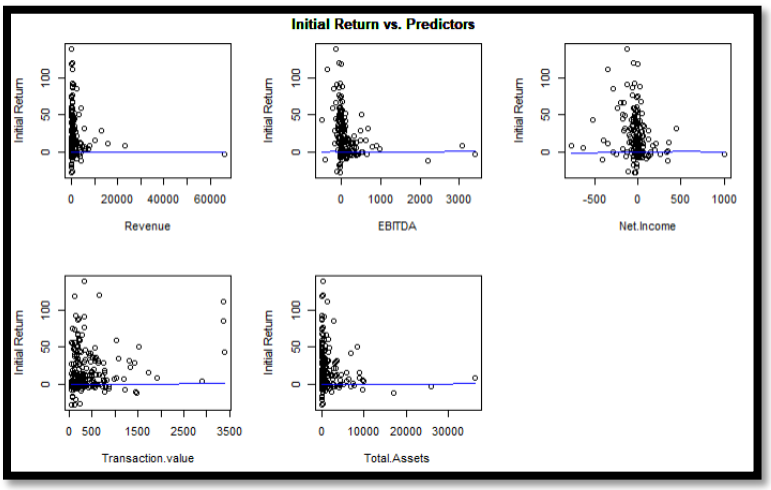
Model summary:

```
Call:
lm(formula = Initial.Return ~ Revenue + EBITDA + Tech + Net.Income +
    Transaction.value + Total.Assets, data = base)

Residuals:
    Min      1Q  Median      3Q     Max
-49.486 -10.468  -5.969   4.195 123.127

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.540e+00  1.784e+00   4.786 2.75e-06 ***
Revenue           9.648e-04  4.682e-04   2.061  0.04023 *
EBITDA           -1.850e-02  1.294e-02  -1.429  0.15401
TechYes           1.939e+01  3.914e+00   4.953 1.26e-06 ***
Net.Income       -1.904e-02  1.549e-02  -1.229  0.22025
Transaction.value  9.715e-03  3.608e-03   2.693  0.00751 **
Total.Assets      -7.004e-05  1.290e-03  -0.054  0.95672
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.57 on 282 degrees of freedom
Multiple R-squared:  0.1762,    Adjusted R-squared:  0.1587
F-statistic: 10.05 on 6 and 282 DF,  p-value: 4.543e-10
```

In multiple linear regression, we assume that the error terms have zero mean, constant variance and that they are independent. That is, the linearity assumption means that the relationship between the response variable and each predicting variable x_j is linear for all predicting variables, the variance of the error terms ε_i is constant across all i = 1,2,3,…,n observations, and that the error terms are independent random variables. We also assume that the error terms are normally distributed for the purpose of statistical inference. In case some of the assumptions do not hold, we interpret that the model fit is inadequate or that the goodness of fit is poor, but it does not necessarily mean that the regression model is not useful.

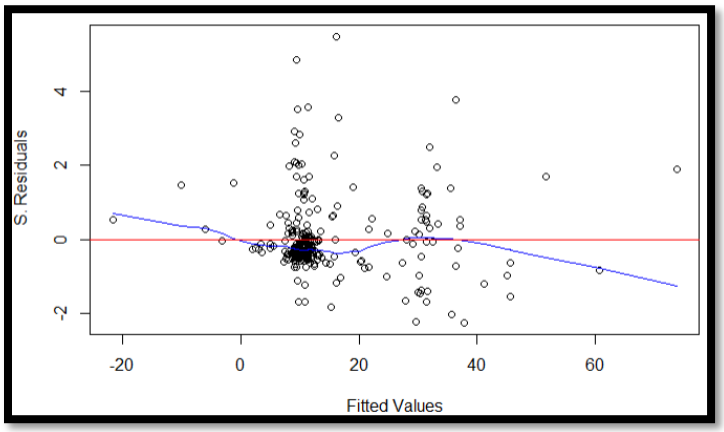
Testing the goodness of fit of the final base model, we plot each x_j predicting variable (other than the categorical variable Tech) against the response and each x_j predicting variable against the standardized residuals to assess the linearity assumption. The base model's standardized residuals mean is zero, but it is difficult to assess the linearity relationship between each predictor and the response since most of the observations are clustered around a particular value.

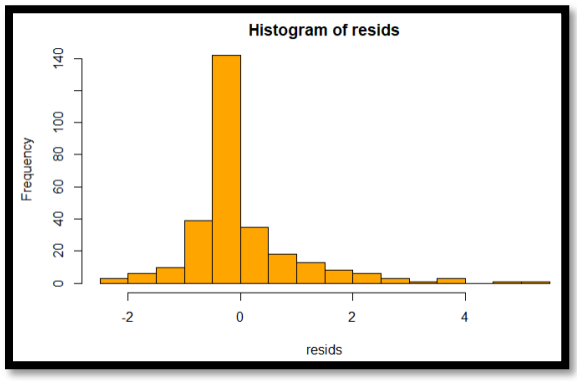


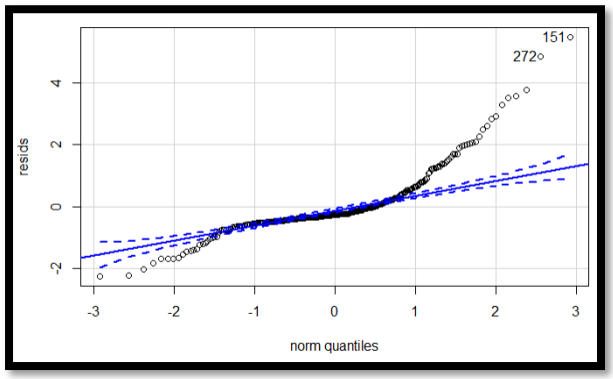


Assessing the constant variance and independence assumptions, we plot the standardized residuals against the fitted values. Since independence is difficult to assess in observational studies (as opposed to randomized trials), we assess the residuals for uncorrelation rather than

independence. Based on the plot of the standardized residuals against the fitted values we conclude that there is no heteroscedasticity (i.e. the variance is constant for the most part) and the residuals are uncorrelated since there are no clusters. However, as indicated by the heavy tailed histogram and QQ plots, the normality assumption does not hold.

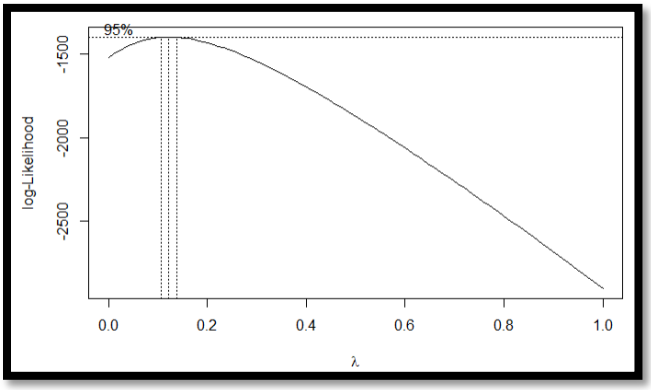






A box-cox transformation performed on the base model suggested a log transformation of the response variable (optimal λ rounded to nearest half integer of 0), but the log model's Adjusted R^2 is low at 0.03654, and has a worse goodness of fit and much larger MSPE and PM compared to the original final base model.

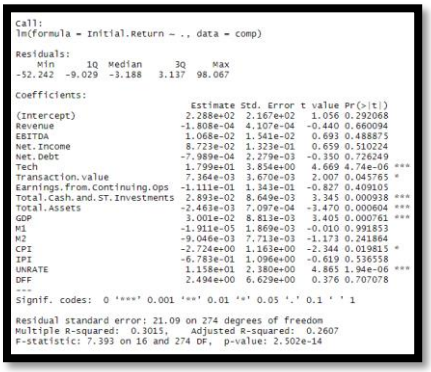
Box-cox optimal λ:



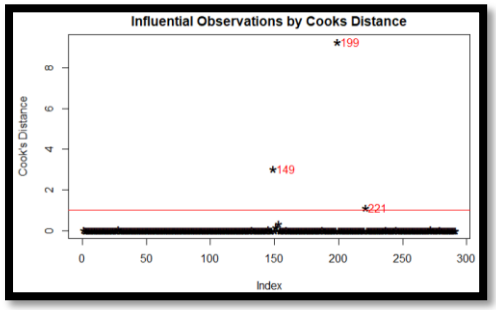
model summary:

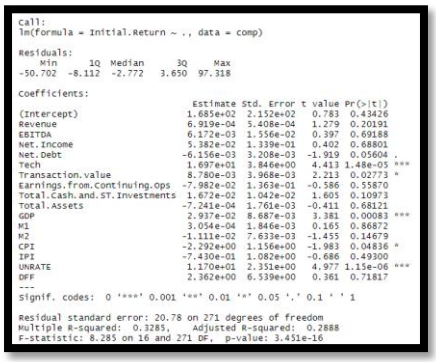
```
Call:
lm(formula = log(Initial.Return) ~ Revenue + EBITDA + Tech +
    Net.Income + Transaction.value + Total.Assets, data = test)

Residuals:
     Min       1Q   Median       3Q      Max
-30.4958   0.2279   5.6099   6.6726  16.4911

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -4.777e+00  9.036e-01  -5.286 2.51e-07 ***
Revenue           3.643e-04  2.359e-04   1.544  0.12361
EBITDA           -7.585e-03  2.878e-03  -2.635  0.00888 **
TechYes           2.390e+00  1.971e+00   1.213  0.22626
Net.Income       -1.332e-02  6.533e-03  -2.040  0.04232 *
Transaction.value  8.782e-04  1.761e-03   0.499  0.61837
Total.Assets      6.451e-05  6.576e-05   0.981  0.32748
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.4 on 282 degrees of freedom
Multiple R-squared:  0.05662,   Adjusted R-squared:  0.03654
F-statistic: 2.821 on 6 and 282 DF,  p-value: 0.01108
```

Fitting the complete model (i.e. with all "firm related" and macroeconomic variables) produces a model with Adjusted R^2 of 0.2607. Fitting the complete model after removing outliers using Cook's distance threshold of 1 produces a model with Adjusted R^2 of 0.2888.



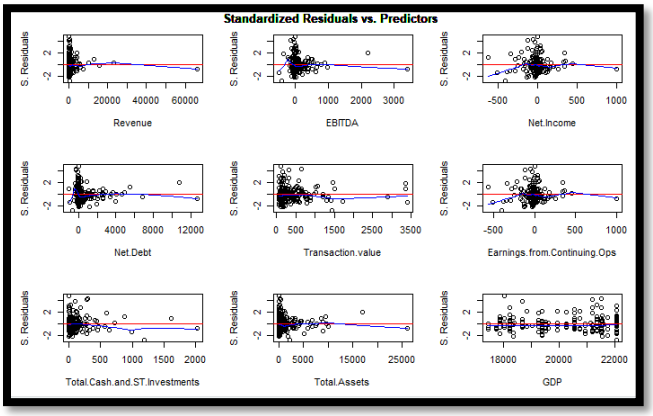
```
Call:
lm(formula = Initial.Return ~ ., data = comp)

Residuals:
    Min      1Q  Median      3Q     Max
-52.242  -9.029  -3.188   3.137  98.067

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.288e+02  2.167e+02   1.056 0.292068
Revenue                     -1.808e-04  4.107e-04  -0.440 0.660094
EBITDA                       1.068e-02  1.541e-02   0.693 0.488875
Net.Income                   8.723e-02  1.323e-01   0.659 0.510224
Net.Debt                    -7.989e-04  2.279e-03  -0.350 0.726249
Tech                         1.799e+01  3.854e+00   4.669 4.74e-06 ***
Transaction.value            7.364e-03  3.670e-03   2.007 0.045765 *
Earnings.from.Continuing.Ops -1.111e-01  1.343e-01  -0.827 0.409105
Total.Cash.and.ST.Investments 2.893e-02  8.649e-03   3.345 0.000938 ***
Total.Assets                -2.463e-03  7.097e-04  -3.470 0.000604 ***
GDP                          3.001e-02  8.813e-03   3.405 0.000761 ***
M1                          -1.911e-05  1.869e-03  -0.010 0.991853
M2                          -9.046e-03  7.713e-03  -1.173 0.241864
CPI                         -2.724e+00  1.163e+00  -2.344 0.019815 *
IPI                         -6.783e-01  1.096e+00  -0.619 0.536558
UNRATE                       1.158e+01  2.380e+00   4.865 1.94e-06 ***
DFF                          2.494e+00  6.629e+00   0.376 0.707078
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.09 on 274 degrees of freedom
Multiple R-squared:  0.3015,    Adjusted R-squared:  0.2607
F-statistic: 7.393 on 16 and 274 DF,  p-value: 2.502e-14
```

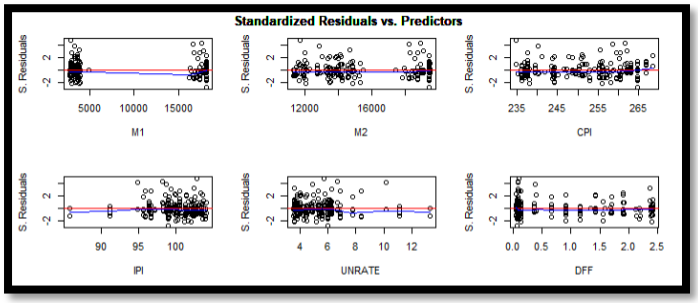


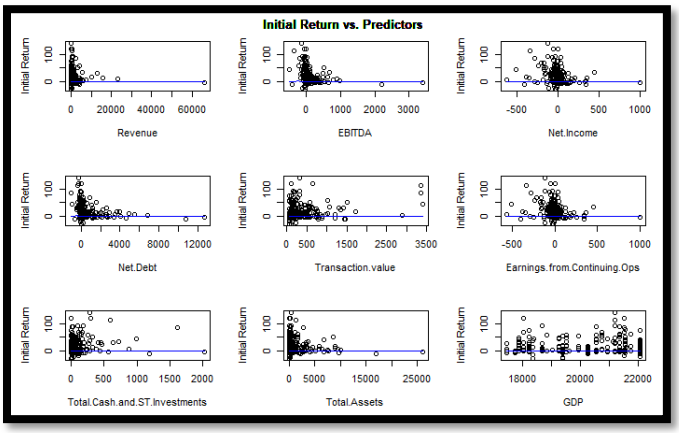


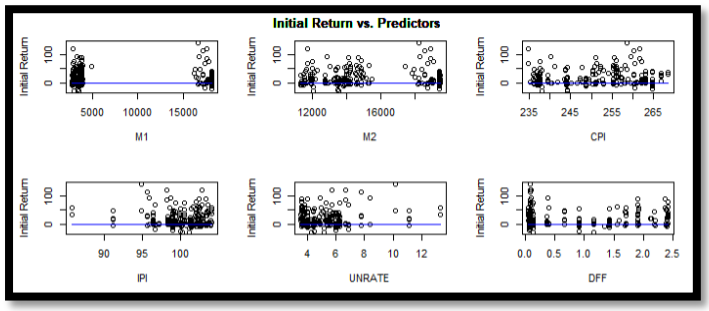
```
Call:
lm(formula = Initial.Return ~ ., data = comp)

Residuals:
    Min      1Q  Median      3Q     Max
-50.702  -8.112  -2.772   3.650  97.318

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.685e+02  2.152e+02   0.783  0.43426
Revenue                      6.919e-04  5.408e-04   1.279  0.20191
EBITDA                       6.172e-03  1.556e-02   0.397  0.69188
Net.Income                   5.382e-02  1.339e-01   0.402  0.68801
Net.Debt                    -6.156e-03  3.208e-03  -1.919  0.05604 .
Tech                         1.697e+01  3.846e+00   4.413 1.48e-05 ***
Transaction.value            8.780e-03  3.968e-03   2.213  0.02773 *
Earnings.from.Continuing.Ops -7.982e-02  1.363e-01  -0.586  0.55870
Total.Cash.and.ST.Investments 1.672e-02  1.042e-02   1.605  0.10973
Total.Assets                -7.241e-04  1.761e-03  -0.411  0.68121
GDP                          2.937e-02  8.687e-03   3.381  0.00083 ***
M1                           3.054e-04  1.846e-03   0.165  0.86872
M2                          -1.111e-02  7.633e-03  -1.455  0.14679
CPI                         -2.292e+00  1.156e+00  -1.983  0.04836 *
IPI                         -7.430e-01  1.082e+00  -0.686  0.49300
UNRATE                       1.170e+01  2.351e+00   4.977 1.15e-06 ***
DFF                          2.362e+00  6.539e+00   0.361  0.71817
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.78 on 271 degrees of freedom
Multiple R-squared:  0.3285,    Adjusted R-squared:  0.2888
F-statistic: 8.285 on 16 and 271 DF,  p-value: 3.451e-16
```

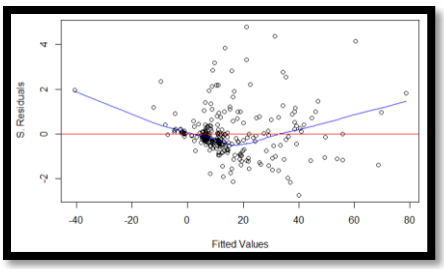
Goodness of fit analysis indicates that the zero mean assumption holds, as well as the constant variance (for the most part) and uncorrelation. However, the normality assumption does not hold. Moreover, as indicated by the corrplot and VIF test, there is a multicollinearity problem.

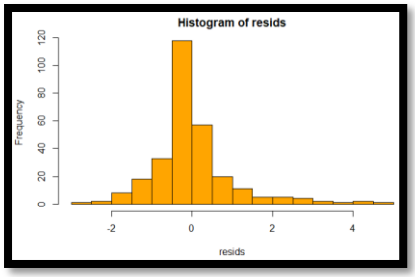


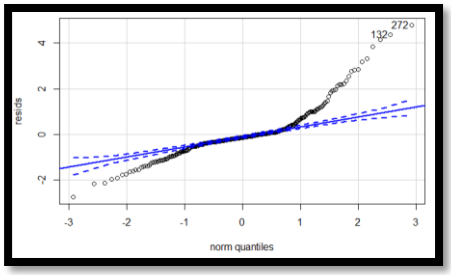


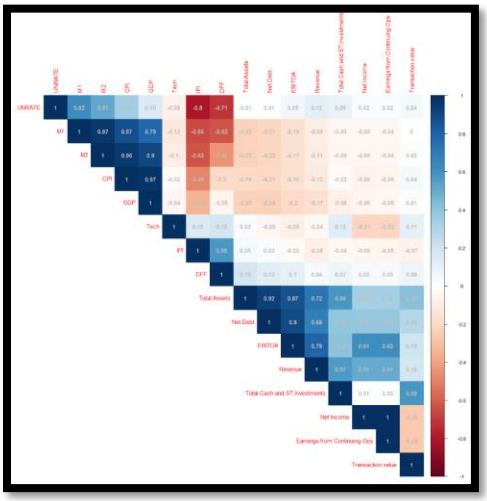












VIF Threshold: 10

Tech	Revenue	EBITDA	Net.Income	Net.Debt
1.204	3.800	13.207	164.948	12.647
GDP	Transaction.value	Earnings.from.Continuing.Ops	Total.Cash.and.ST.Investments	Total.Assets
115.602	2.063	170.666	3.137	13.319
UNRATE	M1	M2	CPI	IPI
7.915	118.899	365.123	89.038	5.963
	DFF			
	16.248			

We used regularization (i.e. Ridge, LASSO and Elastic Net) to address the multicollinearity problem in the complete model. Ridge regression is not variables selection since the coefficients are "shrinking" and none of the them equals 0. Ridge regression is a parsimonious model that performs L2 regularization. The L2 regularization adds a penalty equivalent to the square of the magnitude of regression coefficients and tries to minimize them (but they will not be 0). Hence, we should expect all coefficients to be included in the Ridge model. The LASSO regression on the other hand is an L1 regularization where we add a penalty equal to the sum of the absolute values of the regression coefficients. Minimizing the penalized least squares using this penalty will force some of the coefficients to be 0. In the case where the number of *p* predictors is greater than the number of *n* observations, LASSO will select at most *n* variables. When there is a group of high correlated variables, LASSO tends to randomly select only one variable from the group as opposed to Ridge where the entire group is selected and the multicollinearity is resolved by shrinking the coefficients. The Elastic Net regression combines both the L1 and L2 penalties and we have the advantages of both LASSO and Ridge regressions. By considering both penalties, the model enforces some of the coefficients to be 0 like LASSO and shrinking the others like Ridge (i.e. encourages group effect).

Minimizing the penalized least squares, we have:

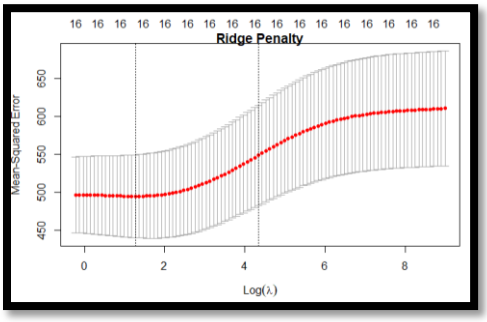
- LASSO: $argmin \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\,\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$
- Ridge: $argmin \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\,\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$
- Elastic Net: $argmin \sum_{i=1}^{n}(y_i - \sum_j x_{ij}\,\beta_j)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2$

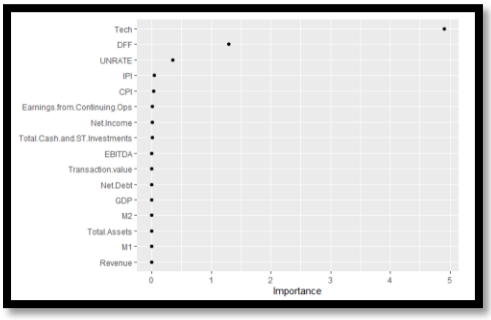
To find the optimal λ for the models above we used cross validation with K-fold = 5 and MSE as measure type. More specifically,

- We divide the data into K = 5 folds
- For a range of λ and for K = 1 to K,
    - Train set consists of data without the *k-th* fold and test set consists of the *k-th* fold.
    - Given λ, fit a model using the train set and predict responses.
    - Compute MSE with the *k-th* fold.
    - Compute overall error (mean MSE) for that λ and all folds.
- Select λ that corresponds to the smallest overall error.

See below MSPE, PM and Adjusted R^2 of the three regularized regressions, Ridge CV and rank of coefficients importance plots as well as LASSO and Elastic Net CV, rank of coefficients and path plots.

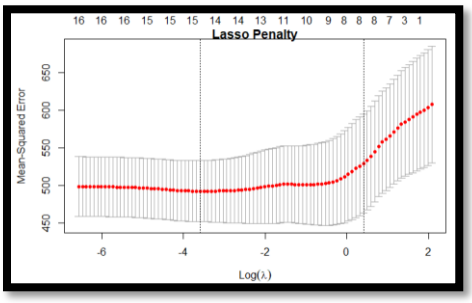
Ridge Regression:
MSPE: 432.5
PM: 0.71
Adjusted R^2: 0.243

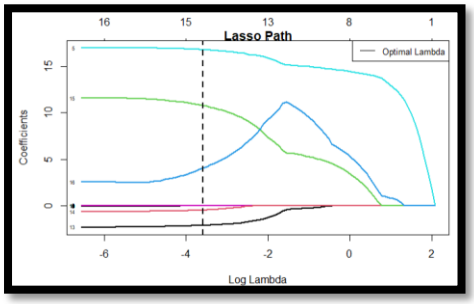




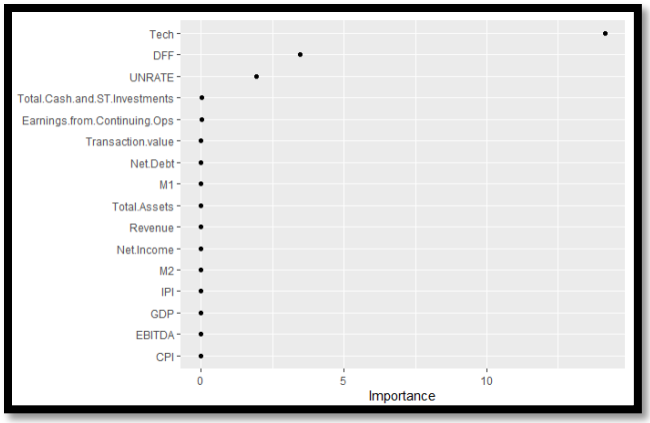

```
17 x 1 sparse Matrix of class "dgCMatrix"
                                          1
(Intercept)                      65.5853245555
Revenue                           0.0004204121
EBITDA                           -0.0060972411
Net.Income                       -0.0063871116
Net.Debt                         -0.0033386761
Tech                             14.4728432767
Transaction.value                 0.0069903575
Earnings.from.Continuing.Ops     -0.0089564008
Total.Cash.and.ST.Investments     0.0147573106
Total.Assets                     -0.0005883498
GDP                               0.0004459726
M1                               -0.0003582637
M2                               -0.0005990088
CPI                              -0.1533022675
IPI                              -0.3701513535
UNRATE                            3.8567571172
DFF                               7.0607824945
```

LASSO:
MSPE: 407.26
PM: 0.67
Adjusted R^2: 0.29

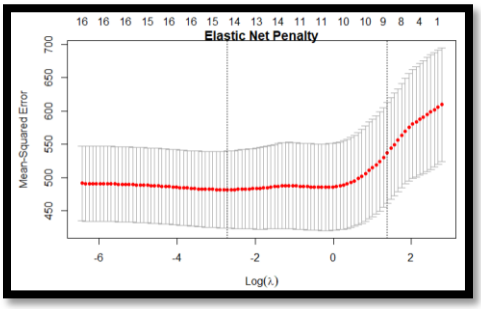


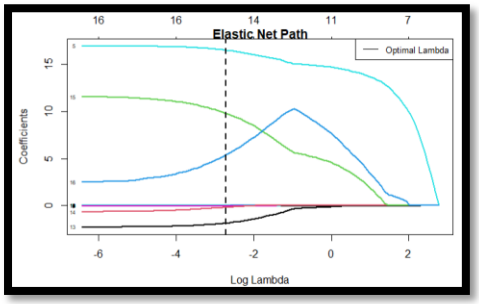



```
17 x 1 sparse Matrix of class "dgCMatrix"
                                          1
(Intercept)                      1.638316e+02
Revenue                          6.905633e-04
EBITDA                           1.410672e-03
Net.Income                       .
Net.Debt                        -5.858916e-03
Tech                             1.678705e+01
Transaction.value                8.746049e-03
Earnings.from.Continuing.Ops    -2.184832e-02
Total.Cash.and.ST.Investments    1.550279e-02
Total.Assets                    -4.535698e-04
GDP                              2.390040e-02
M1                              -6.876537e-06
M2                              -7.945916e-03
CPI                             -2.110138e+00
IPI                             -4.613703e-01
UNRATE                           1.077798e+01
DFF                              4.073967e+00
```

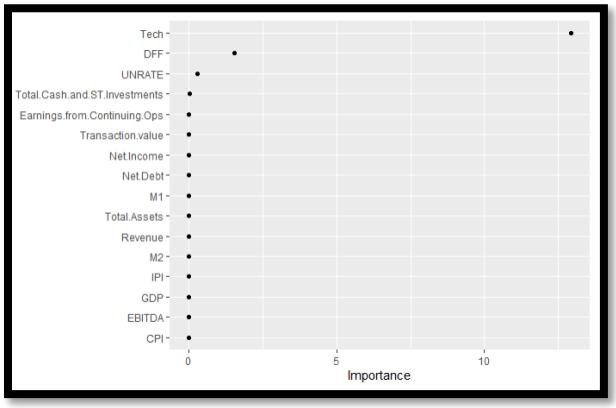


Elastic Net:
MSPE: 409.23
PM: 0.68
Adjusted R^2: 0.29






```
17 x 1 sparse Matrix of class "dgCMatrix"
                                        1
(Intercept)                  1.427985e+02
Revenue                      6.652933e-04
EBITDA                       .
Net.Income                   .
Net.Debt                    -5.708915e-03
Tech                         1.651590e+01
Transaction.value            8.757110e-03
Earnings.from.Continuing.Ops -1.992257e-02
Total.Cash.and.ST.Investments 1.490994e-02
Total.Assets                 -4.188584e-04
GDP                          1.819369e-02
M1                           -4.873057e-04
M2                           -4.586655e-03
CPI                          -1.847060e+00
IPI                          -1.839117e-01
UNRATE                       9.821457e+00
DFF                          5.401493e+00
```



Results summary and Conclusion

Predicting IPOs initial return has been attempted for many years. However, previous work has focused solely on "firm related" variables. This paper has taken the assumption that together with "firm related" variables, macroeconomic indicators can increase the prediction accuracy. As the summary table below indicates, all three complete regularized models performed better than the base model in terms of Adjusted R^2 and prediction accuracy. The complete Ridge model's Adjusted R^2 has increased by almost 10%, an increase of more than 50% compared to the base model, and the complete LASSO and Elastic Net models have

increased by almost 15%, an increase of close to 100%. The complete LASSO model performed the best in terms of prediction accuracy with the lowest MSPE and PM. The MSPE has decreased by 18% compared to the base model's MSPE. Also, in terms of coefficients importance, we observed in the "importance" plots that the most important coefficients include two of the macroeconomic indicators (i.e. DFF and UNRATE) and one "firm related" variable (i.e. Tech). Moreover, the coefficients that have been forced to 0 in both the LASSO and Elastic Net models are from the "firm related" variables (i.e. EBITDA and Net Income) , suggesting that the macroeconomic indicators contain more predictive power.

Model/ Indicator	**Base**	**Complete Ridge**	**Complete LASSO**	**Complete Elastic Net**
Adjusted R^2	15.87%	24.3%	29%	29%
MSPE	496.88	432.5	407.26	409.23
PM	0.824	0.71	0.67	0.68
Model Rank	4	3	1	2

In this paper we showed that including macroeconomic variables could significantly improve IPOs initial return prediction when compared to "traditional" methodologies. Hence, we suggest that future work should include such predictors regardless of the algorithms used (e.g. regression, ANN, etc.) as it is expected to generate a more accurate prediction.

Reference:

[1] Anirban Sen. (2021, June 21). *U.S. IPOs hit annual record in less than six months*. https://www.reuters.com/business/finance/us-ipos-hit-annual-record-less-than-six-months-2021-06-15/

[2] Baba Boubekeur & Sevil Guven. Predicting IPO initial returns using random forest. *Borsa Instanbul Review.* 2020; 20(1): 13-23. https://www.sciencedirect.com/science/article/pii/S2214845019302686#bbib51

[3] Hanley, K.W. The underpricing of initial public offerings and the partial adjustment phenomenon. *J. Financ. Econ*. 1993; 34, 231–250. https://www.sciencedirect.com/science/article/pii/0304405X93900198

[4] Huang et al. A genetic-search model for first-day returns using IPO fundamentals. *International conference on machine learning and cybernetics, IEEE, Xian*. 2012; 1662-1667. https://ieeexplore.ieee.org/abstract/document/6359624

[5] Jamaani Foud & Alidarous Manal. Review of Theoretical Explanations of IPO Underpricing. *Journal of Accounting, Business and Finance Research*. 2019; 6(1): 1-18. https://www.researchgate.net/publication/335919142_Review_of_Theoretical_Explanations_of_IPO_Underpricing

[6] Loughran, T., Ritter, J.R. & Rydqvist, K. Initial public offerings: International insights. *Pacific-Basin Finance Journal*, 1994; 2(2-3): 165-199. Updated March 22, 2021. https://site.warrington.ufl.edu/ritter/files/International.pdf

[7] Luque et al. Predicting IPO underpricing with genetic algorithms. *International Journal of Artificial Intelligence*. 2012; 8 (S12): 133-146. http://davidquintana.apps-1and1.net/wp-content/uploads/2017/11/Predicting-IPO-Underpricing-with-Genetic-Algorithms.pdf

[8] Mitsdorffer Rolf & Diederich Joachim. Prediction of first-day returns of initial public offering in the US stock market using extraction from support vector machines. *Rule extraction from support vector machines* 2008; Vol. 80, Springer: 185-203. https://link.springer.com/chapter/10.1007/978-3-540-75390-2_8

[9] PriceWaterhouseCoopers. (2021, May 20). *The SPAC spree: Current state.* https://viewpoint.pwc.com/dt/us/en/pwc/in_the_loop/in_the_loop_US/the_rise_of_SPACs.html

[10] Quintana D, Sáez Y, Isasi P. Random Forest Prediction of IPO Underpricing. *Applied Sciences*. 2017;7(6):636. https://doi.org/10.3390/app7060636

[11] Reber et al. Predicting mispricing of initial public offerings. *Intelligent Systems in Accounting, Finance and Management*. 2005; 13 (1): 41-59. https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.253

[12] Ritter JR. The costs of going public. *Journal of Financial Economics*. 1987;19(2):269-281. https://www.sciencedirect.com/science/article/pii/0304405X87900055

[13] Robertson et al. Neural network models for initial public offerings. *Neurocomputing*. 1998; 18 (3): 165-182. https://www.sciencedirect.com/science/article/pii/S0925231297000775

[14] University of Florida, Warrington College of Business. IPO Data. https://site.warrington.ufl.edu/ritter/files/IPOs-Underpricing.pdf

[15] Wang et al. An interpretable neural fuzzy inference for predictions of underpricing in initial public offerings. *Neurocomputing*. 2018; Vol. 30: 102-117. https://reader.elsevier.com/reader/sd/pii/S0925231218308713?token=BCB8D81F21EA7C2E6602A0B9232F8EA3C8F8452CD4342599FA73EDC3A821D9AF7D1F8D1819F231584BFEFE03A1D23774&originRegion=us-east-1&originCreation=20210622004328

[16] Weng Bin et al. Macroeconomic indicators alone can predict the monthly closing price of major U.S. indices: Insights from artificial intelligence, time-series analysis and hybrid models. *Applied Soft Computing*. 2018; 71: 685:697. https://www.sciencedirect.com/science/article/pii/S1568494618304125#bib0155