EXECUTIVE SUMMARY OF NALCHIK INVESTMENTS PROJECT

The Nalchik Investments collected data from Bank of America, and I as a data analytics professional of data team should to provide data-driven suggestions based on the understanding of the data to the stakeholders. They want to know which sectors of the US economy are the best in terms of the total value of company assets and other indicators in order to choose at this stage the direction of investment inflows into the economy.

Company has the following question:

1. Which sectors of US market are the best for investing?
2. Which sectors have the biggest sum of market caps and which sectors are the most promising?
3. Which economic parameters (from this dataset) are the most important for the investing decision and what other data should be collected?

Goals in this project are to analyze the dataset, understand which sectors of US market are the best for investing and to build a model that **predicts stock prices of companies**.

About dataset

The dataset was taken from Bank of America.

The dataset provides a comprehensive view of various companies that are categorized under the sector, industry and originate from the USA. It includes details such as the 'Ticker', 'Company', 'Sector', 'Industry', 'Country', 'Market Cap', etc. (other columns are shown in Jupiter Notebooks)

However, the dataset also includes companies not regarded as USA companies.

PLAN STAGE

EDA

GOALS OF THE PROJECT

The Nalchik Investments wants data specialists to identify the sectors of the economy that occupy the largest share and influence the market using the current dataset for the period specified in it.

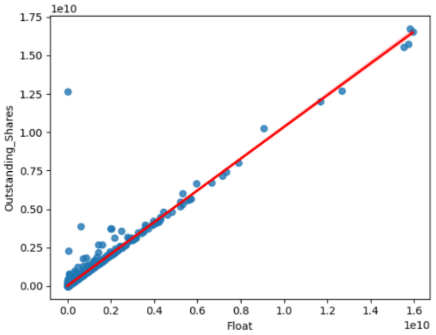
And the second goal for the data team (me as a part of the team) is to develop a machine learning model to predict the stock price of companies in the US stock market using different economic parameters. To begin, the data team needs to organize the raw dataset and prepare it for future exploratory data analysis.

IMPACT

According to the findings from the exploratory data analysis, the future stock price model will need to account for null values and imbalance almost in all columns by incorporating them into the model parameters.

DETAILS ABOUT THE DATASET

1. many columns of the data frame have missing values. The dataset consisted of 63 columns, and 29 of them had more than 3000 missing data;
2. after descriptive statistics, it became obvious that in almost 80% of the columns there are outliers;
3. duplicates of columns were found, namely Volume. One of them was removed from the data frame after checking for correlation.
4. Preparation for the EDA revealed a clear correlation between 'Float' and 'Outstanding Shares'.



NEXT STEPS

1. all of missing values should be cleaned up, some columns should be completely removed from the dataset due to the inability to fill in the data;
2. all missing data should not be thrown away because there were 9501 rows in the data frame, and after cleaning it will be less than 1500 rows, which is a loss of more than 80% of the rows;
3. during analyzing, duplicates will need to be correctly processed, either thrown out of the data frame or replaced;
4. duplicate of the column 'Volume' should be removed from the data frame after checking for correlation;
5. using linear regression, I will fill in the missing data in the 'Outstanding Shares' column before analyze stage, because of correlation with 'Float', and moreover it's only 72 rows, so it won't be such a big influence on the model.

ANALYZE STAGE

STATISTICS AND VISUALIZATIONS

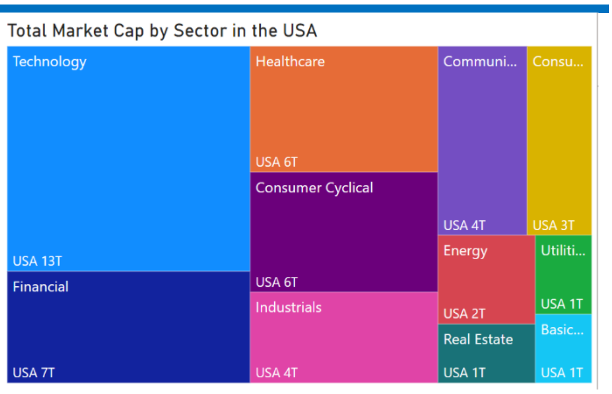
PROJECT STATUS

Data has been cleaned and prepared for EDA.

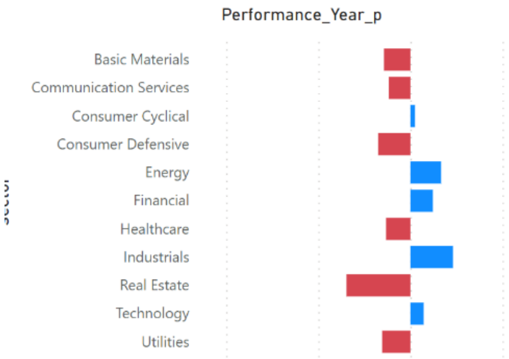
GOALS OF THE STAGE

1. to answer the question -which sectors have the biggest sum of market caps and which sectors are the most promising.
2. Find out what correlations there are between various variables (especially correlations with stock prices).
3. If necessary, normalize data for subsequent construction of a predictive model

KEY INSIGHTS



1. The sectors of the US economy that occupy the largest share are technology, financial companies, medicine and the consumer sector, obviously in terms of market capitalization in the same order, 13 trillion dollars of total capitalization - the technology sector, 7 trillion dollars - the financial sector.
2. The highest return on assets and capital among energy companies, followed immediately by financial companies, while the rest of the companies from the top-4 sectors, oddly enough, occupy the last places in terms of profitability.



3. The correlation heatmap show that 'Price' has huge positive correlation only with each 'Average_True_Range', 'Performance' columns have positive correlation with 'Simple_Moving_Average' columns, 'Yearly_High' and 'Yearky_Low' column, and 'Relative_Strength_Index'. And whether 'Weekly_Low_p' is negatively correlated with 'Monthly_Low_p'
4. Now let's move directly to the share price. Financial companies are again the most stable, as can be seen from weekly, monthly, quarterly and annual indicators. If we look at the long term (that is, for a period of a year or more), then again financial, technology and consumer sector companies show the best economic results, and also, oddly enough, industrialists.

NEXT STEPS

Create a predictive stock price model and evaluate the accuracy of the model. Find out what economic parameters are key and suggest what other data is needed for the model.

CONSTRUCT STAGE

MACHINE LEARNING MODEL OUTCOMES

GOAL OF THE STAGE

Nalchik Investments wants to buy stocks of US companies, and data team should understand how prices can change, so data team should create a predictive stock price model and evaluate the accuracy of the model.

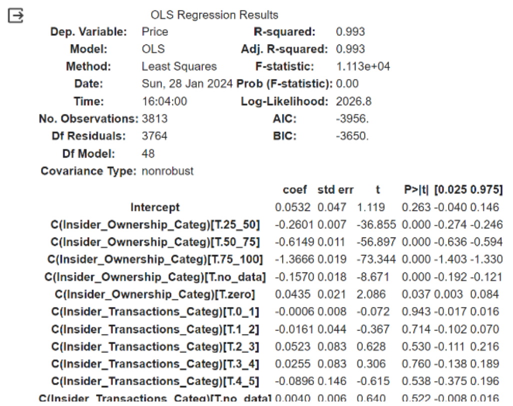
SOLUTION

The data team built two Linear Regression models. Both models were used to predict on a held-out validation dataset, and final model was determined by the model with the best R^2. The final model was then used to score a test dataset to estimate future performance.

DETAILS

Both model architectures — Linear Regression (LR) and OLS — performed well, but the OLS model had a better R^2 and MSE and was selected as champion. Performance on the test holdout data yielded near perfect scores.

Subsequent analysis indicated that the primary predictors were all related to ‘Price_to_Sales’, ‘Price_to_Cash’, ‘Float’, ‘Short_Ratio’, ‘Beta’ and other. This conclusion was made on the basis that their n-values are less than 0.05.



```
                          OLS Regression Results
      Dep. Variable:  Price                      R-squared:  0.993
             Model:  OLS                   Adj. R-squared:  0.993
            Method:  Least Squares            F-statistic:  1.113e+04
              Date:  Sun, 28 Jan 2024 Prob (F-statistic):  0.00
              Time:  16:04:00              Log-Likelihood:  2026.8
 No. Observations:  3813                              AIC:  -3956.
      Df Residuals:  3764                              BIC:  -3650.
          Df Model:  48
  Covariance Type:  nonrobust
                                                coef    std err  t        P>|t|  [0.025  0.975]
                                  Intercept     0.0532  0.047    1.119    0.263  -0.040  0.146
      C(Insider_Ownership_Categ)[T.25_50]      -0.2601  0.007   -36.855   0.000  -0.274  -0.246
      C(Insider_Ownership_Categ)[T.50_75]      -0.6149  0.011   -56.897   0.000  -0.636  -0.594
     C(Insider_Ownership_Categ)[T.75_100]      -1.3666  0.019   -73.344   0.000  -1.403  -1.330
    C(Insider_Ownership_Categ)[T.no_data]      -0.1570  0.018   -8.671    0.000  -0.192  -0.121
       C(Insider_Ownership_Categ)[T.zero]       0.0435  0.021    2.086    0.037   0.003   0.084
    C(Insider_Transactions_Categ)[T.0_1]       -0.0006  0.008   -0.072    0.943  -0.017  0.016
    C(Insider_Transactions_Categ)[T.1_2]       -0.0161  0.044   -0.367    0.714  -0.102  0.070
    C(Insider_Transactions_Categ)[T.2_3]        0.0523  0.083    0.628    0.530  -0.111  0.216
    C(Insider_Transactions_Categ)[T.3_4]        0.0255  0.083    0.306    0.760  -0.138  0.189
    C(Insider_Transactions_Categ)[T.4_5]       -0.0896  0.146   -0.615    0.538  -0.375  0.196
C(Insider_Transactions_Categ)[T.no_data]        0.0040  0.006    0.640    0.522  -0.008  0.016
```

NEXT STEPS

As noted, the model performed well on the test holdout data. Before deploying the model, the data team recommends further evaluation using additional subsets of company data. Furthermore, the data team recommends monitoring the distributions of ‘Price’ and ‘Market Cap’ levels to ensure that the model remains robust to fluctuations in its most predictive features.

EXECUTE STAGE

PROJECT GOALS

Goals in this project are to analyze the dataset, understand which sectors of US market are the best for investing and to build a model that **predicts stock prices of companies**.

PROJECT KEY INSIGHTS

1. many columns of the data frame have missing values. The dataset consisted of 63 columns, and 29 of them had more than 3000 missing data and all of missing values were cleaned up, some columns were completely removed from the dataset due to the inability to fill in the data
2. The sectors of the US economy that occupy the largest share are technology, financial companies, medicine and the consumer sector, obviously in terms of market capitalization in the same order, 13 trillion dollars of total capitalization - the technology sector, 7 trillion dollars - the financial sector and in the future Nalchik Investments should try to predict stock prices from such sectors,
3. The highest return on assets and capital among energy companies, followed immediately by financial companies, while the rest of the companies from the top-4 sectors, oddly enough, occupy the last places in terms of profitability.
4. Financial companies are the most stable, as can be seen from weekly, monthly, quarterly and annual indicators. If we look at the long term (that is, for a period of a year or more), then again financial, technology and consumer sector companies show the best economic results, and also, oddly enough, industrialists.
5. OLS prediction model had a better R^2 and MSE and was selected as champion. Performance on the test holdout data yielded near perfect scores.

PROJECT REPORT

The model performed well on the test holdout data. Before deploying the model, the data team recommends further evaluation using additional subsets of company data.

NEXT STEPTS

Furthermore, the data team recommends monitoring the distributions of ‘Price’ and ‘Market Cap’ levels to ensure that the model remains robust to fluctuations in its most predictive features.