**Introduction to Machine Learning Report**

**Group-04**

Benjamin Jeffrey - 14-921-530

Carlo Scherrer - 13-930-821

Alexander Wyss - 15-736-374

# Introduction

In the context of the seminar “Introduction to Machine Learning” we were assigned with the task to create a machine leaning program written in Python to predict stock movements of 30 companies in the Dow Jones Industrial Index. To do so we trained and tested our algorithms on 10 years of data starting from Jan-2006 to Dec-2015. Our algorithms classify movements of the selected stocks in up and down movement. The strength of the movement is therefore not predictable with our model.

In the following paper we describe which methods and data we used to zero in on the best binary classification model for our 30 stocks in the given period.

# Data Preparation

The primary source of data used in this project, was downloaded from the Wharton database (2018a, 2018b) in two separate Csv-files. One contains data on the stock market and the other on various ratios of the companies. To obtain these we went through all the available features to download from the database and selected the ones we thought could be significant to predict stock movement. These two datasets were merged using Excel before loading into the Python environment. This was done because they didn't have the same number of rows and consequently the dates didn't match up. By deleting some duplicate lines, we got everything lined up correctly.

All further cleaning and adjusting of the data was done within Python by utilising the Pandas library to manipulate our data-frame. First, we analysed where a lot of NaN values were occurring and deleted features with too much missing data. With smaller patches of missing data, where we could justify filling with generated values, we analysed the data points individually and decided on the best method to fill the missing data. We used a combination of filling using backwards-fill, forwards-fill, filling with the mean of the feature, and filling with the mean of the company in question.

With no missing values left in the data frame, we identified features that contained categorical data. We then created Dummy features for each category of those features so that the algorithms could differentiate the categories through numeric values. We also wanted to include a categorical feature to show which industry the company belongs to. The categorization from the database was so fine however, that every company had its own category which was a useless feature to us. So, we created our own categories and added them to the Csv-file.

The last step to prepare our data for use with the algorithms was to create a response vector from the price column. Basically, if the stock price rose the next month the response vector contains a 1 and for downwards motion a 0. After removing the price column from the data-frame we had our data fully ready to be fed into the algorithms. The train-test splits were done as preparation for each algorithm, as we dropped some features for Logistic regression for instance which necessitated a different split.

# Logistic Regression

Logistic Regression was easy enough to implement, the hard part was trying to increase the rather disappointing test score. We tried data normalization, data standardization and adjusting for class imbalance. All of the above actually decreased our performance metrics. We then looked into dropping some of the features using recursive feature scaling. It selects the most relevant features by recursively considering smaller and smaller feature sets to find the optimal set. Utilising this we managed to improve our performance metrics to a point where we felt we could not get more out of this algorithm with our knowledge. We achieved a train score of 59.95% and a test score of 58.96% with our standard train-test split of 80:20.

# Support Vector Machine

We started by creating a linear svm object and applied to our data. The object presents a linear svm with C=0. To expand our model, we added the non-linear Kernel functions poly, sigmoid and rbf. The results we got:

* Linear: train score of 62.23% and a test score of 59.24%
* poly: train score of 67.55% and a test score of 56.44%
* sigmoid: train score of 52.17% and a test score of 54.76%
* rbf: train score of 67.55% and a test score of 57.56%

Because of possible class imbalances we added the class weight in the svm creation function and inserted the value “balanced”. We applied this to all non-linear Kernel functions. We discovered that all the test score decreased about 1% to 4%.

To improve our model, we added the pipeline function and grid search to discover the best combination of values for C and polynomial degrees. The best result we get for C is 10 and the best polynomial degree is one, which means nothing else than a linear function. As you can see above, the best result, comparing all the four Kernel functions, we had with the linear svm object.

# Decision Tree

For the Decision Tree model, we also selected all our features and first tried manually to find the best parameters. But quickly we decided to use our computing power at hand to evaluate the ideal constellation of parameters for our Decision tree model. Because our model contains 83 features (including dummy variable for 30 companies) we decided that our tree should be allowed to have a reasonably large size.

Using GridSearch we iterated through the following parameters:

* criterion: Define the function to measure the quality of a split: ‘gini’ or ‘entropy’
* max\_depth: specifies the maximum depth of the tree
* min\_samples\_split: Minimum number of samples required to split an internal node
* min\_samples\_leaf: Minimum number of samples for each split.

Not to our surprise the best parameters included a depth of our tree of 8. Instantly the suspicion arouse that we have overfitted our model, but with a train score of 76.42% and a test score of 70.03% we decided that this is a reasonable result.

# Random Forest

The random forest model was in the beginning one of the most intriguing to us, because we assumed that we will have many correlating features. So, the idea to grow ‘n’ decision trees from the bootstrap sample and randomly select a defined number of features and aggregate the ‘n’ predictions and assign the class label by majority vote looked very promising.

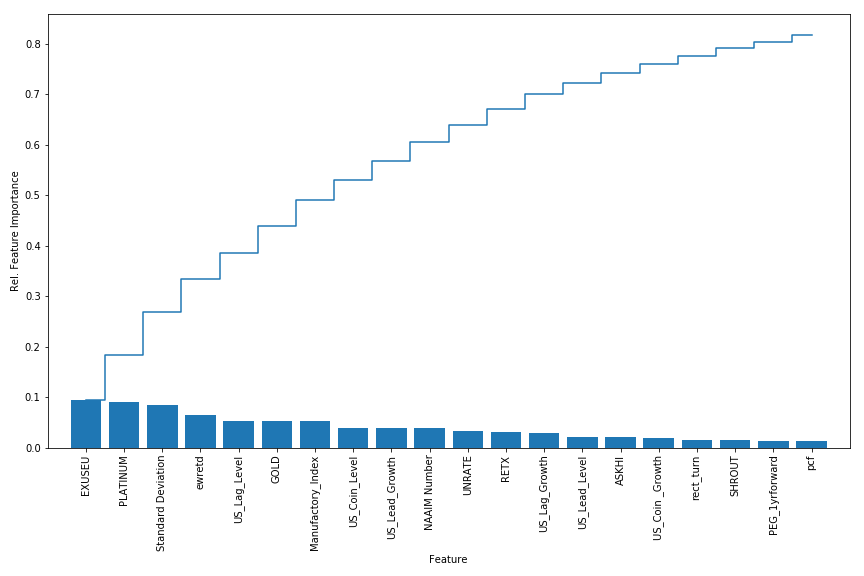
In the first try we used the same parameters we found in our decision tree grid search: But the result was a sobering 65.27%. Because this result was far from our expected outcome, we decided to implement a separate grid search for the RandomForestClassifier. Besides the parameters already specified in the decision tree grid search, we added the following:

* n\_estimators: Number of trees in the forest
* max\_features: The number of features to consider when looking for the best split
* verbose: Controls the verbosity of the tree building process
* warm\_start: Defines if it should reuse the solution of the previous call to fit and add more estimators to the ensemble or if a whole new forest should be fit

The end result was 76.21% train score and 71.85% test score.

# Feature Importance

After we applied all the methods, we wanted to know, which feature of our dataset explains our model the best. We extracted the feature labels and created a Random Forest object, fit the data and extracted the feature importance attributes. To have a better overview we plotted our top twenty features.



It was nice to see, that ten of the best features were added additionally, so we made a good selection.

# Explanation of additionally added features

US Civilian Unemployment Rate, UNRATE (FRED (2018a)):

We have chosen to add this feature because The Federal Bank interest rates influence the labour market, but also the entire market. Especially in the US this interaction can be observed very well (diekleinanleger.com (2014)).

US / Euro Foreign Exchange Rate, ESUSEU (FRED (2018b)):

Currency exchange rates affect travel, exports, imports and the economy. Furthermore, our dataset includes only US companies and the biggest part of them exports a lot to Europe (Investopedia (2018)).

National Association of Active Investment Managers, NAAIM Exposure Index and National Association of Active Investment Managers Standard Deviation, NAIIM Standard Deviation (Quandl (2018d)):

The NAAIM Exposure Index represents the average exposure to US Equity markets reported by NAAIM members.

The NAAIM Exposure Index was designed to present the progression of change in adviser sentiment. Furthermore, it contains only the US Equity Market and because of that it is a very good feature to compare with our dataset (NAAIM (2018)).

Gold Price, GOLD (Quandl (2018c)):

Theoretically there is an inverse relationship between the stock market and gold prices. There have been circumstances where the stock markets rise and gold prices fall. Gold prices may also rise in sympathy with the fall in stock prices. The reason lies in the perception of the market by investors. Investors who foresee a bearish market, usually take positions in gold futures to safe guard their investments. Furthermore, a lot of the companies in the data set are part of the technology industry, where gold is often used to produce electronical goods (ResearchGate (2018)). Increasing gold prices are a traditional indicator of a recession or a downturn in an economy. People run to the safety of gold when they think the value of other investments may go down in the future (Quora (2014)).

Moreover, two of the most important stock markets for gold are located in the US, the New York Mercantile Exchange and the Chicago Board of Trade (finanzen.ch (2018)).

Platinum Price, PLATINUM (Quandl (2018e)):

As well as gold, platinum is a raw material often used in the technology and electronic goods industry. Prized for its extreme resistance to corrosion and its ability to catalyse reactions, such as those vital in the refining of crude oil into gasoline. Platinum is essential in a variety of laboratory and industrial applications. The catalytic converter, a vehicle emissions control device, is the most widely used application of platinum group metals. Platinum is also used for biomedical applications, crucibles, dental alloys, electronic devices, glass, hard disks, silicones, turbine blades and watches (Investopedia (2018)).

Empire State Manufacturing Index, Manufactory Index (Quandl (2018b)):

The fact, that the index represents the economical healthiness of the US manufactory industry and that most of the companies in the dataset have their headquarters in New York, it is a suitable feature explaining a hole industry (Investing (2018)).

US Coincident Index, US\_Coin\_Level, US\_Coin\_Growth (Quandl, 2018a):

Because the index contains some interesting and important economic measures, it could be very decisive for our dataset. Furthermore, the trend for each state’s index is set to the trend of its GDP (Trading Economics (2018)).

US Lagging Index, US\_Lag\_Level, US\_Lag\_Growth (Quandl (2018f)):

This lagging indicator follows an economic event. We used it to confirm what recently happened in the economy. In general lagging indicators are statistics and are especially useful for identify turning points in the business cycle (the balance (2017)).

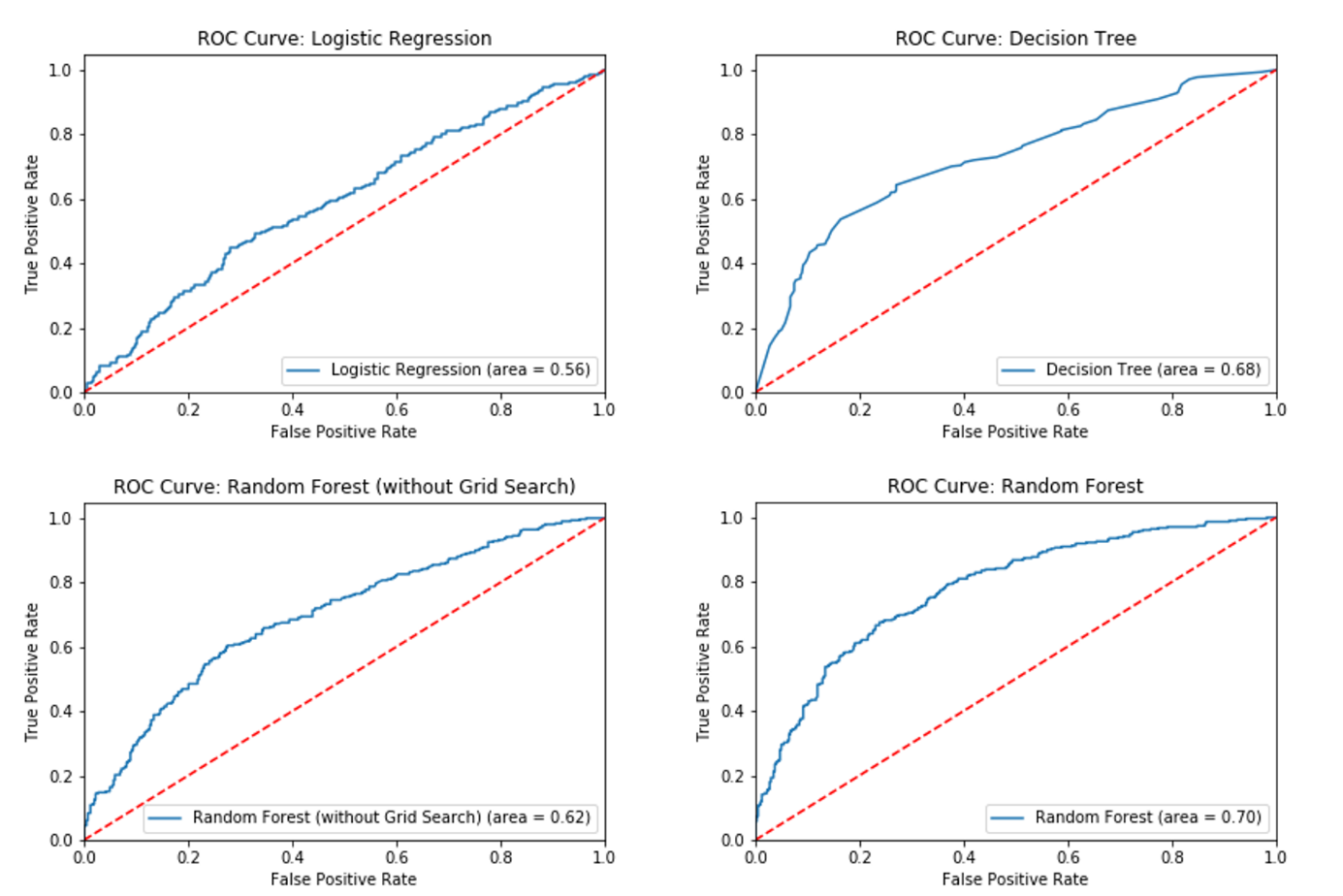
US Leading Index, US\_Lead\_Level, US\_Lead\_Growth (Quandl (2018f)):

The fact, that the Leading Index is just one part of an array of leading indexes of US economic growth which, taken together, make up a formidable defence against surprise cyclical upturns and downturns in economic growth. In other words, the index contains data which correlate with cycling points it could have a coherency with our stock prices (ECRI (2018)).

# Results

Comparing the results of our models many of our outcomes were to be expected. Since we analysed a complex Problem with many features our Logistic Regression and QDA did not show very good results, even after reducing the number of features to only the most relevant.

On the other hand, we expected our best results in SVM or Decision Tree/Random Forest. To our surprise SVM scored with 57.56% worse than Logistic Regression (58.54%). The fact that the best polynomial degree in our SVM model was to the power of one explains partially why the Logistic Regression scored so good in comparison to our SVM model.



By far the best results were reached in our Decision Tree and Random Forest models as it can be seen in our ROC graphs. With a Test score of 71.85% Our Random Forest with Grid search is our best result.

# Experiences

After finalising our work, we reflected on our developing process. For us, the biggest challenges were handling missing data and finding the right parameters for our models. Furthermore, we applied our model just to one stock and the results were much better. However, after further investigation we found other companies were far worse so we refrained from conducting theses test for each company individually. Moreover, we had to be aware, that we do not add too much additional features and bias our model. One bias we have for sure is the look-ahead bias, because of applying forwards fill (Investopedia (2018)).

On the other hand, there were some steps worked very well. The whole data import process we managed very well as well as merging the two data files together. Furthermore, the implementation of the y-vector and applying the different machine learning algorithm to our dataset we managed pretty well. Moreover, we did a great additional data selection, which improved our results a lot, as you can see in the feature importance section.

Overall, we are happy with our results and the hole project was also fun. As a team we worked very well together. Our meetings were always very efficient and the different tasks and work load were always clear and divided fairly.

# Bibliography

Amadeo, Kimberly, 2017, What Lagging Economic Indicators Tell You from website, https://www.thebalance.com/lagging-economic-indicators-list-index-and-top-3-3305860, 08.04.2018.

Curtis, Glenn, 2018, The Impact Of Currency Conversions from website, https://www.investopedia.com/articles/basics/07/currencies.asp, 09.04.2018.

ECRI, 2018, The Best Way to Use the Weekly Leading Index (WLI) from website, https://www.businesscycle.com/ecri-reports-indexes/report-summary-details/economic-cycle-research-institute-ecri-lakshman-achuthan-the-best-way-to-use-the-weekly-leading-index, 08.04.2018.

Finanzen.ch, 2018, Goldpreis from website, https://www.finanzen.ch/rohstoffe/goldpreis, 09.04.2018.

Folger, Jean, 2018, Commodities: Platinum from website, https://www.investopedia.com/university/commodities/commodities15.asp, 09.04.2018.

FRED, 2018a, Civilian Unemployment Rate from website, https://fred.stlouisfed.org/series/UNRATE, 24.03.2018.

FRED, 2018b, U.S. / Euro Foreign Exchange Rate from website, https://fred.stlouisfed.org/series/EXUSEU, 24.03.2018.

Investing.com, 2018, USA - NY Empire State Herstellungsindex from website, https://de.investing.com/economic-calendar/ny-empire-state-manufacturing-index-323, 08.04.2018.

Investopedia, 2018, Look-Ahead Bias from website, https://www.investopedia.com/terms/l/lookaheadbias.asp, 11.04.2018.

Mewati, Sunny, 2014, What is the relation between Oil prices, Gold, Stock Markets and Dollar prices? From website, https://www.quora.com/What-is-the-relation-between-Oil-prices-Gold-Stock-Markets-and-Dollar-prices, 09.04.2018.

NAAIM, 2018, NAAIM Exposure Index from website, https://www.naaim.org/programs/naaim-exposure-index/, 08.04.2018.

Onoh, John, 2018, What is the relationship between Gold Price and Stock Market? from website, https://www.researchgate.net/post/What\_is\_the\_relation\_between\_Gold\_Price\_and\_Stock\_Market, 09.04.2018.

Quandl, 2018a, U.S. Coincident Index from website, https://www.quandl.com/data/ECRI/USCOIN-U-S-Coincident-Index, 20.03.2018.

Quandl, 2018b, Empire State Manufacturing Index - Forecasted Technology Spending - Diffusion Index (Not Seasonally Adjusted) from website, https://www.quandl.com/data/FRBNY/TSFDINA-Empire-State-Manufacturing-Index-Forecasted-Technology-Spending-Diffusion-Index-Not-Seasonally-Adjusted, 18.03.2018.

Quandl, 2018c, Gold Price: Fixing from the website, https://www.quandl.com/data/LBMA-London-Bullion-Market-Association, 26.03.2018.

Quandl, 2018d, NAAIM Exposure Inedx from website, https://www.quandl.com/data/NAAIM/NAAIM-NAAIM-Exposure-Index, 26.03.2018.

Quandl, 2018e, Platinum Fixing from website, https://www.quandl.com/data/LPPM/PLAT-Platinum-Fixing, 25.03.2018.

Quandl, 2018f, U.S. Lagging Index from website, https://www.quandl.com/data/ECRI/USLAGGING-U-S-Lagging-Index, 20.03.2018.

Teufl, Andreas, 2014, Arbeitslosenzahlen und Aktienmarkt: Die Wechselwirkung from website, https://www.diekleinanleger.com/arbeitslosenzahlen-und-aktienmarkt-die-wechselwirkung/, 09.04.2018.

Trading Economics, 2018, United States Coincident Index from website, https://tradingeconomics.com/united-states/leading-economic-index, 08.04.2018.

Wharton, 2018a, CRSP Monthly Stock, 12.03.2018.

Wharton, 2018b, Financial Ratios Firm Level by WRDS (Beta), 12.03.2018.