Abdullah Salau

Senior Project Proposal

Stockify: Model for Estimating Stock Performance

9/2/2021

Table of Contents

Summary………………………………………… Page 3

Significance……………………………………… Page 3

Required Tools and Availability………………… Page 4

Demonstration Plans…………………………….. Page 4

Qualifications……………………………………. Page 5

Project Specification…………………………….. Page 5

Technical Details………………………………... Page 6

Timeline…………………………………………. Page 6

Future Enhancements……………………………. Page 9

**Summary**

Stockify is a program containing estimation models designed to predict whether or not a particular stock is worth buying (ie. its value will increase) at the end of a year based on numerous financial indicators that act as the model’s independent variables. The program includes multiple binary classification algorithms for prediction, but also showcases algorithm performance results to determine which model works best with an existing dataset that includes stock information from 2014-2018.

Stockify also contains models to predict one of the nominal financial indicators included in the dataset. Similar to the binary classification problem, Stockify also includes the algorithm performance results for each model designed to predict the nominal financial indicator to determine the highest performing algorithm.

**Significance**

Majority, if not all, of this project is based on the material covered in the Data Mining course, which deals with machine learning and regression analysis. The Python programming language will be used for the extensive preprocessing of the dataset. This includes data cleaning techniques such as one-hot encoding, regularization, handling missing values, and the conversion of categorical and ordinal variables into numeric attributes. Other advanced preprocessing techniques will also be implemented to the dataset, such as feature engineering and Principal Component analysis. This project will also include the creation of classifiers and models that were taught in the Data Mining class, such as Logistic Regression and K Nearest Neighbor, but will also contain numerous other prediction algorithms for both binary and continuous dependent variables, such as Ridge regression, Lasso regression, and Gradient Tree Boosting.

LaTex and IguanaTex will be utilized to present technical information, such as mathematical formulas, in an organized and elegant way. The reports created will contain various visualization elements like plots charts and graphs with the help of Matplotlib to create a better understanding of the dataset.

**Required Tools and Availability**

The classifiers in this project will be created using the Spyder Python 3.8 IDE, which is included in the Anaconda Individual Edition. This version of Anaconda is available for free download. The additional libraries and packages that will be utilized for the models and visualization elements include Pandas, Scikit-learn, NumPy and Matplotlib, all of which can be easily installed via the Spyder IDE. To create LaTex reports, I will utilize Overleaf, which is a free online LaTex Editor.

**Demonstration Plans**

For each checkpoint meeting, I will deliver a PowerPoint presentation which will detail my progress with creating my prediction models. Some of the presentations will include visualization elements created with Matplot lib, LaTex equations using IguanaTex, LaTex reports using Overleaf, and screenshots of chunks of code that show how I accomplish very important aspects of my project. I will be presenting in the Julian Science Center in room 260, where I will utilize the moderator computer to demonstrate experiments focused on scoring numerous algorithms to test model efficiency on the Spyder IDE. I will use a HDMI connector to present my laptop screen.

**Qualifications**

Majority of the machine learning skills and knowledge I possess stem from the Data Mining course I took in the Spring semester of 2021. The class introduced me to the Python Programming language and made me proficient in employing data science procedures using add on libraries and packages such as Pandas, Numpy, and Scikit Learn. I have had basic experience with using LaTex in my Statistical Computation class, but the Data Mining class made me even more qualified by introducing me to Overleaf and helping me learn how to create full LaTex reports.

**Project Specification**

Stockify’s main task is to use prediction models to predict whether a specific stock should be bought or not next year, due to the rise (or fall) of the stock’s value. The program will also be able to estimate one of the nominal financial indicators included in the dataset for each stock listed. My plan to apply for a PhD in Business Analytics immediately after graduation is my main motivation for this project, and also because it combines relevant information from both of my majors: Computer Science and Actuarial Science.

The functional specification of this project lies mainly with its ability to provide fairly accurate predictions. This will strongly be determined by algorithm scorers, such as K-fold Cross Validation, which grade how effective a specific algorithm is in prediction or estimation. Other secondary functionalities include the creation of LaTex reports containing visualization elements to make aspects of the program easier to understand.

**Technical Details**

Extensive preprocessing will need to be done to the dataset in order for algorithms to be able to train itself on it and make predictions. Firstly, models and algorithms can only work with numerical data, so the preprocessing must include steps taken to convert all non-numeric data, such as categorical data and ordinal data, into numeric data. Secondly, preprocessing will include steps to fill in missing values, of which there are numerous strategies to accomplish. Once there are no missing values and all the data is numeric, the algorithm can train itself on the dataset; however, more preprocessing and tuning techniques will need to take place to ensure the model makes fairly accurate predictions. Hyperparameter tuning procedures will also take place to ensure the model architecture is optimal. To test how accurate a model is, Stockify will primarily employ algorithm scorers, such as K-fold Cross Validation. Multiple preprocessing techniques can be employed, and Stockify shows how each technique contributes to the effectiveness of the model. There will also be numerous models and classifiers to try out, each depending on the kind of output to be predicted. Python makes it very easy to test out these models with the Scikit-learn library, which includes packages specially designed for different models. Stockify documents the results of these experiments using LaTex reports and visualization elements which can be formed using the very useful Python package known as Matplotlib.

**Timeline**

Checkpoint 1:

* I will demonstrate the results of basic cleaning of the dataset by showing 5 rows of the dataset before and after the tuning techniques have been implemented. This subset of the dataset will show that all missing values of the dataset have been taken care of, basic one-hot encoding has been implemented, and all categorical and ordinal variables have been converted to numerical variables.
* I will conduct experiments demonstrating the effectiveness of a simple logistic model using a K-folds Cross Validation scoring system on the dataset before and after basic preprocessing has been accomplished to show the effect of the tuning techniques aforementioned.
* I will use IguanaTex to include the formulas I utilized in creating my model inside my PowerPoint Presentation.
* As a guide for feature selection, I will use Matplotlib to create charts that show collinearity between different independent attributes.

Checkpoint 2:

* I will demonstrate the results of more advanced preprocessing by showcasing effects of newly engineered features, feature selection, regularization, etc. I will then estimate model performance on newly processed dataset by using logistic regression and compare it to the model performance of the dataset specified in Checkpoint 1.
* I will conduct scoring experiments with at least 3 Binary Classification algorithms: Logistic Regression, Gradient Tree Boosting, and K-Nearest Neighbor.
* I will include plots and diagrams using Matplotlib in my presentation to further exhibit estimator performance under different hyperparameter values to make it more understandable.

Checkpoint 3:

* A new nominal dependent variable will be selected from the existing dataset. (**price var)** I will demonstrate results of additional preprocessing, such us updated feature selection and feature scaling. I will then test this dataset on a simple linear regression model.
* I will also evaluate the effectiveness of the simple linear regression model on the dataset before it was transformed to show that the new cleaning techniques make the algorithm more efficient.
* I will demonstrate the results of my hyperparameter optimization experiments. This will include at least 2 hyperparameter tuning techniques namely Grid search and Bayesian optimization.
* I will create a LaTex report using Overleaf detailing cross-sector performance for 5 years. The report will be 2 pages long detailing performance trends for each sector using visualization elements such as graphs, charts, and plots.

Checkpoint 4:

* I will demonstrate the performance of at least 3 regression models namely Multiple Linear Regression, Ridge Regression, and Lasso Regression. Additional machine learning algorithms which are yet to be determined will also be included in my experiments.
* I will use Matplotlib to create charts that will show regression performance under different hyperparameter values, after which I will compare between the highest performing algorithm and the regression done in Checkpoint 3 to show improvement.
* I will document all my experiments and trials using Overleaf and come up with a final report that will also contain details about features included in the final processed dataset.

**Future Enhancements**

* Implement a User Interface. This will allow users to input their own data e.g entering financial indicator data for a stock they are interested in buying. Stockify will then predict if the stock’s value will rise or fall the following year.
* Implementing more preprocessing techniques and more classifiers/algorithms. This will give the program more opportunities to create more accurate prediction models.
* Expanding the dataset to include financial indicator data from years prior to 2014 and years following 2018.

**Bibliography**

“200+ Financial Indicators of US Stocks (2014–2018).” *Kaggle*, 18 Jan. 2020, www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018.

“Pre-Modeling: Data Preprocessing and Feature Exploration in Python.” *YouTube*, uploaded by Next Day Video, 10 May 2016, www.youtube.com/watch?v=V0u6bxQOUJ8.