sddec18-13: Asset management - Financial Factor Discovery - "Value"

Week 6 Report
March 1 - March 7

Team Members

Carter Scheve — *Communications Lead*
Nathan Hanson — *Project Progress Tracker/Manager*
Caleb Utesch — *Meeting Scribe*
Jack Murphy — *Research Analyst*
Samuel Howard — *Lead Engineer*
Alex Mortimer — *Project Manager*

Weekly Summary

This is the sixth weekly report for our team. The rate of progress has been rapid lately, which has led to more challenging issues as we confront more difficult tasks. We feel as though we are past the basic implementation stages of our predictive models, and moving forward into the more advanced implementation details that pose more issues. As we finish with basic parts of our project, we have taken some time to pause on the development track and take time to test what we have created so far. This involves both the accuracy of our models and the data we're using to train them, as well as the train-test split that has been headed by Carter the last couple of weeks.

Summary of Progress this Report

- Wrote our own train-test data splitting method
    - Ensured that there are no future data being used in the training steps
    - Provided time buffer equivalent to the amount of time into the future user is trying to predict
    - Will be largely useful for the remainder of the project
- Tested data importing library
    - Provided a set of fresh eyes that have no experience with or knowledge of how the tool works
    - Improved the robustness of the instructions
    - Included instructions for different operating systems, programs used, etc
- Eliminated k-Nearest Neighbors and Support Vector Machines as predictive models
    - May still use for factor selection

Pending Issues

- Still no concrete technique to prove model accuracy
- Signal level in the data is miniscule, so finding trends is difficult
- Client has requested research into additional technologies with which we have little or no experience
- Train-test library both needs more functionality and more robustness

---

Plans for Upcoming Reporting Period

- Continue testing changes in the out-of-box models for accuracy and speed
- Tune parameters for models to gauge improvements
- Begin comparing models to finalize a model to implement moving forward

Individual Contributions

Team Member	Contribution	Weekly Hours	Total Hours
Carter Scheve	This week, was a lot of time spent in the data formatting, gathering, and testing category. I first started creating a train_test_split algorithm that we, as a group and client decided on in our last meeting. I also fixed the data that we use in our models as it was skewed. I updated the library with a new functionality to categorize data better. I also updated documentation for the data library. After that was complete, I spent some time re running models with our new mechanism to see what our preliminary results were.	10	60
Nathan Hanson	Spent a few hours researching methods for batch jobs for feature selection. Readied my algorithm for use with batch job (even though we are throwing SVM out - i did this just for practice and proof of concept). Managed repository structure, uploaded necessary documents to website. Began exploring methods for efficient feature selection.	11	57
Caleb Utesch	The majority of my time this week was spent looking into several different feature selection methods. I tried out several different methods including univariate selection, SelectFromModel selection, and L1-based selection. The only one I was able to get to work was the L1-based feature selection. I think the others should be able to work after putting some more time in and figuring out how to format the data before importing it. Will start looking into tree-based feature selection starting the week after spring break as it was recommended by our client	11	52
Jack Murphy	Looked into Bernoulli Bayes classifier machine learning algorithm. Worked on trying to improve the train_test_split that we were currently using. Did some research into feature selection to improve the prediction of our models.	10	54

Samuel Howard	This week we all spent a lot of time ensuring that our train-test data splits were correct and not including any future-dependant data in the predictions. Tried experimenting with different parameters, and found out that the ideal AR value was 2. However, since we are predicting a year out, and the model requires making estimates on every intervening week and then uses those predictions, we are dropping the autoregressive models for return prediction. We have instead been directed to use the models on the features to decorrelate and/or make the data stationary.	11	55
Alex Mortimer	This week we all spent a lot of time ensuring that our train-test data splits were correct and not including any future-dependant data in the predictions. I also began the process of testing Carter's data importing library, and gave feedback to make it more robust and easy to follow for users on different operating systems, with no experience in the language used, and other things of that nature. I will continue that process in the coming weeks as he iterates through improvements in the library. As far as my model, I will begin researching rolling window data selection to improve results.	11	63