**Weekly Update**

**Oct 12 - Oct 16**

This week, our group focused on generating a list of proposed input variables for our model. Additionally, we set up a meeting with Cary to discuss these variables and other assumptions we are considering for our models. Lastly, we began to compile information on the dependent variable in terms of stock prices for companies where job hiring information was collected by Greenwich. What follows is our initial work on this subject.

**Columns should be used in each table:**

Master: *job\_id, company, post\_date, salary* (Not sure about geographic data, seems irrelevant)

Role: *job\_id, role*

Tag: *job\_id, tag* (not sure about this yet)

Timelog: *job\_id, post\_date, remove\_date, duration\_days*

Titles: *job\_id, title*

One assumption is that we should treat an individual company in a specific timeframe as a single entry of training data. For example, a company with one year of data can be treated as 12 entries of training data for a monthly stock timeframe.

**List of proposed x input variables:**

* **Job posting change(in percentage):** The increase in active job postings for a company in month N in comparison to month N-1. Active job postings for each month can be inferred from *remove\_date* and *post\_date* in Timelog.
* **Average salary change(in percentage):** The change in average salary level for a company in month N in comparison to month N-1. Average salary for each month can be calculated from the *salary* column in Master.
* **Weight of important roles change(in percentage):** The change of  weight of important roles for a company in month N in comparison to month N-1. Important roles can be defined as roles such as managers, VP, Chief...Officer, etc. We can use regular expressions to identify those important roles for the *role* column in Role.
* **Average job posting duration days:** The average duration days for jobs that are removed in month N for a company. We can use *duration\_days* and *remove\_date* in Timelog.
* **Weight of intern change(in percentage):** The change of  weight of intern roles for a company in month N in comparison to month N-1. Interns can be identified using regular expression by searching for “intern” or “internship” or “trainee”, etc.

**Y variable:**

* **Stock price change(in percentage):** The change in stock price in month N+1. Calculated from stock price at the last trading day in month N+1 and that in month N. Use this for our regression model.
* **Stock price change(+/-):** Use this for our classification model.

**Training model:**

* **Multiple Linear Regression**
* **SVM**

Initially, we plan to use the SVM model to predict the trend of the stock while using the multiple linear regression to predict the percentage change in the stock price.

**Cross Validation:**

* **Use K-fold:** We will divide historical data for each company into two groups, one for training and one for testing.

**Accuracy Evaluation:**

* **Criteria for “accurate”:** If the actual increase in stock price is larger than or equal to the predicted increase in stock price or the absolute value of actual decrease in stock price is larger than or equal to the absolute value of predicted decrease in stock price, then we can call it an accurate prediction.
* **F-Score:** Calculate precision, recall and F-score.

Next week, Congda and Binqi will be in charge of manipulating the job data and transforming the columns to quantitative measures(e.g. percentage change in job posting, average salary change, percentage of important roles) to better fit the model. In addition, Matt and Isaac will be responsible for fetching historical financial data based on the company symbol column, which is normalized and cleaned this week. Last but not the least, we will start building our model with the input vectors that are confirmed with Cary.