# CSE 482 FINAL PROJECT (Cover Page)

Project Title:

Predicting Future S&P 500 from Economic Indicators using LSTM

Summary of Team Member Participation:

Fill out the following table for each team member of the group.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Name** | **Participate in data collection** | **Participate in preprocessing** | **Participate in data analysis/ experiment** | **Participate in writing the final report** | **Participate in creating video presentation** | **Completed Assigned Tasks** |
| Andreas Frame | x | x | x | x | n/a | yes |
| Jon Spiwak | x | x | x | x | n/a | yes |

Team Member Roles and Contributions:

|  |  |
| --- | --- |
| **Name** | **Roles and Contributions** |
| Andreas Frame | Data collection, primary data preprocessing. Helped with data analysis and experimenting. Primary writing of final report. |
| Jon Spiwak | Data collection and preprocessing. Primary role in data analysis and experimenting. Helped writing the final report. |

I approve the content of the final report (please add your signature below):

Andreas Frame: *Andreas Frame*

Jon Spiwak: --------------------------------------------------

**Predicting Future S&P 500 from Economic Indicators using LSTM**

Andreas Frame, Jon Spiwak

Project URL: <https://github.com/akframe123/bigdataproject>

**ABSTRACT**

*Write a short abstract that summarizes what the project is about. Explain briefly the motivation behind the project and state the specific goals you would like to achieve from the project. Briefly summarize the approach you use (e.g., I formulated the problem as a binary classification or regression or clustering problem and applied method X to solve it). Finally, discuss the findings of the project and state whether you were able to achieve the goal*.

The S&P 500 Index, hereafter referred to as the S&P, is a stock market index measuring the performance of the top 500 companies listed on the US stock exchange. As such, it is an excellent indicator of the stock market as a whole, and is heavily influenced by economic indicators. The project aimed to use indicators such as interest rates, the DFF and VIX in order to train a Long short-term memory (LSTM) model to predict the future S&P. Using this approach, we were successful in building a model to accurately predict the next days S&P given previous data.

# INTRODUCTION

For introduction, you need to include the information given below. Use this as a guide to organize your writing. Try to provide as much detail as possible (e.g., each bullet below could be expanded to 1 or 2 paragraphs):

1. Start by providing some background description of the problem. Specifically, explain the application domain of the project and give a rationale why this is an important or interesting problem. Provide other background information needed to understand the application domain. For example, if the project is about predicting the outcome of an online game between two teams, explain what that game is about, what is the scoring system, is there only 1 or multiple players per team, etc. If the project is about sentiment analysis on twitter data for some predictive modeling application, explain what sentiment analysis is and why it is useful for the problem.

The stock market is a massively complex system dependent on a multitude of factors. Despite this, the stock market directly and indirectly affects peoples lives around the country. Particularly in recent times, the volatility and unpredictability of the stock market has been on full display. Consequently, building a model that can predict what the S&P will be when the market next opens is extremely beneficial to traders and many others. As such, modeling and predicting the S&P was a significant undertaking. This was ultimately performed using a LSTM, a specific kind of recurrent neural network designed with the ability to learn long-term dependencies. This was especially useful in this project, as it allowed the model to form dependencies between events over large periods of time.

1. State the goal or hypothesis of the project. For example, the goal/hypothesis might be using machine learning models can help to make accurate prediction or reliable recommendation in the given application domain; or that sentiment analysis from twitter data can provide useful features to accomplish certain task.

This project aimed to accurately predict the next days S&P before the market opens using just historical data from the S&P and other key economic indicators. Economic indicators found to be highly correlated with the S&P

1. Explain clearly how you plan to achieve the goal. For example, if the goal is to determine whether user sentiment in twitter data provides useful information to predict changes in the price of a stock, then the plan could be to apply regression method on a combination of historical price and twitter sentiment data along with other stock price indicators to improve the prediction.
2. Briefly describe how do you conduct the project, e.g., what type of data have you collected or analyzed in order to test the hypothesis (you don’t need to give too much details about the data here as you will provide the in-depth discussion about the data in Section 2).
3. Explain the challenges encountered when collecting, preprocessing or analyzing the data. Describe what are the steps you took to overcome these challenges.
4. Discuss briefly the findings of the project. What were the results and were you able to achieve the goal or prove/disprove the hypothesis.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*CSE881-2015*, Month 1–2, 2004, City, State, Country.

Copyright 2004 ACM 1-58113-000-0/00/0004…$5.00.

# DATA

Give a detailed explanation about the datasets you use. Specify the sources of the raw data (give a URL citation), how you collect them, and what is the underlying format (e.g., CSV, JSON, etc). Did you use a crawler or write some script to collect the data or are you simply downloading the data (e.g., from Kaggle website). Do you have to fuse/merge data from different files and if so, how did you do it (e.g., did you join them based on some key attributes). What are the features/attributes in the raw data you have collected (before preprocessing). Provide a table that summarizes the list of raw features/attributes in the original data. If there are too many features, you can list a few important ones that you end up using in the project.

|  |  |  |
| --- | --- | --- |
| **Attribute name** | **Type** | **Description** |
| Timestamp | Ordinal | Date and time of event |
| Measurement | Ratio | Measurement of event |
| … | … | … |

**Table 1**: Attributes of the data acquired from XXX (data source).

Next, discuss the characteristics of the data you have collected. For example, if it is a time series data (e.g., tweets or stock market data), how long is the time period (e.g., 2 years of data from April 2016 to April 2017). If it is spatial data (e.g., crime data), what is the location or spatial coverage (e.g., crime for New York city or Chicago). Explain some of the data issues that must addressed. For example, does it have a lot of irrelevant features that must be discarded? Does it have a lot of missing values? What is the size of the raw data (how many sample observations and features)? Summarize the information in a table.

|  |  |
| --- | --- |
| Number of observations | 300,000 (~100 MB) |
| Number of attributes | 300 time steps (Apr 2016-Apr 2017) |
| % missing values | 2.4% |
| … | … |

**Table 1**: Summary statistics of the raw data from XXX (source).

State the preprocessing steps you have performed on the raw data. For example, did you have to impute the missing values or did you simply discard them? Did you have to merge data from multiple sources? Did you create a sample of the raw data, and if so, what is your sampling approach? Did you do any Z-score transformation to remove trends or seasonality in the data? Did you find any outliers and remove them? Did you apply any discretization method on the attributes (e.g., to convert time of event to time of day – morning, afternoon, evening)? If so, what method did you use? Did you have to perform some calculations to create the data frame object you need to apply classification or regression? For example, to predict the outcome of a game, did you have to calculate some statistics about the teams based on their previous k games and use them as features to predict the outcome of their next game?

Finally list the set of features you use to create the Dataframe object for your data analysis software. For example, if it is a classification or regression problem, explain what are the predictor attributes used and the corresponding response (target) attribute. If it is a classification problem, state what are the classes. If it is a recommender system problem, explain how the final data look like (e.g., is it in a 3-column format: user, product, rating). How many users and products are there? If it is a crime data, how many neighborhoods and type of crimes are there?

If possible, list all the attributes you have created in a Table (similar to Table 1). Report the size of the final dataset after preprocessing, i.e., the number of rows and columns of the data frame you have created. The final dataset corresponds to the data frame object will be provide to the data analysis software (e.g., scikit learn or Python surprise) for model fitting/training. You should also report the dataset size in Mbytes. You should upload the final dataset you have created on the project GitHub page (unless the size is too large). Do not upload the raw data unless you have permission from the original data owner. In particular, Twitter prohibits any publication of their raw data on public domains. Kaggle also has similar policies. Make sure you read carefully the acceptable use policy of the website from which you have downloaded the data. If you have any questions about publishing the final (processed) data, please inform the instructor. If it is human subject data, make sure there are no identifiable attributes in the published data.

Note that this is an important section to demonstrate to the instructor the amount of work you did for the project. If all you did was to simply download a dataset that has already been nicely prepared (e.g., Kaggle data) and there is little evidence you had performed much preprocessing, then you will not receive as much credit for the project.

# METHODOLOGY

This section should present details of the methodology you have implemented for the project. If possible, draw a high-level diagram or flowchart to summarize the process – from data collection and preprocessing to data analysis. State clearly what method you used to do the analysis (e.g., I use support vector machine or decision tree classifier to perform my classification or I use matrix factorization from Python Surprise toolkit to do the recommendation).

If it is a predictive modeling or recommendation problem, describe how you create the training and test sets. Did you do cross-validation to select the model hyperparameters? If it is a clustering problem, what is the distance measure you use? How do you decide the number of clusters in the data?

Finally, give a brief summary of the the code you have written for this project. For example, you can summarize it as follows:

* Collection.ipynb: this is the Jupyter notebook file that I wrote to collect data from Twitter.
* Preprocess.ipynb: this is the Jupyter notebook file to look for tweets that contain a specific keyword and do sentiment analysis on the tweets. The output of the script is ….
* Modeling.ipynb: this is the Jupyter notebook file to perform the classification task of the project.

State what other auxiliary software you need to replicate the results of your project. For example, if you’re using some sentiment analysis software, make sure you report it in this section. Give the URL of the software.

# EXPERIMENTAL EVALUATION

This section describes the experimental setup and results you obtain.

## Experimental Setup

This section should include:

1. Computing platform (what operating system and hardware you use to do the experiment). Are you using AWS cluster? If so, how many nodes?
2. What baseline methods did you to compare against your approach? Make sure you report the results of your method as well as the baseline methods. This is important to demonstrate whether your project was successful.
3. What evaluation metric did you use to report the results (e.g., accuracy, root-mean-square-error, F1-measure, etc).

## Experimental Results

This section should include a description about the experiments you have performed and the results obtained. If possible, try to summarize the results in a figure or a table. In addition to reporting the final performance numbers (e.g., accuracy), a good project report should examine in details the model created by the data analysis software. For example, if you’re building a regression or classification model, check the model coefficients to determine what features are important for making the model predictions (see the lecture notes on how to access the model coefficients). If you’re doing clustering with k-means, how do the centroid vectors look like? What are the important features characterizing each cluster centroid? Discuss the significance of the results or any new unexpected insights revealed by the results.

Finally, state whether the project was successful. If not successful, explain the reasons that could affect the poor results you have obtained. What steps you could have taken to alleviate the problem.

The grading of the project does not only depend on how much have you accomplished in this project, but also the level of detailed analysis you performed to make sure your findings are correct.

# CONCLUSIONS

Summarize the overall findings and contributions of the project. If possible, provide suggestions for future work that could improve what you’ve done for the project.

# REFERENCES (at least 3 references)

1. Bowman, B., Debray, S. K., and Peterson, L. L. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst., 15,* 5 (Nov. 1993), 795-825.
2. Ding, W., and Marchionini, G. *A Study on Video Browsing Strategies.* Technical Report UMIACS-TR-97-40, University of Maryland, College Park, MD, 1997.
3. Fröhlich, B. and Plate, J. The cubic mouse: a new device for three-dimensional iput. In *Proceedings of the SIGCHI conference on Human factors in computing systems   
   (CHI ’00)* (The Hague, The Netherlands, April 1-6, 2000). ACM Press, New York, NY, 2000, 526-531.
4. Lamport, L. *LaTeX User’s Guide and Document Reference Manual.* Addison-Wesley, Reading, MA, 1986.
5. Sannella, M. J. *Constraint Satisfaction and Debugging for Interactive User Interfaces.* Ph.D. Thesis, University of Washington, Seattle, WA, 1994.

Grading criteria

Note that the project accounts for 10% of your final grade. The project will be graded based on the following criteria:

1. Presentation - structure/organization and clarity of writing (including tables and figures).
2. Technical - Correctness and thoroughness of the analysis performed. What are the challenges faced and how well did you address them? How do you evaluate the performance of the method you'd applied to the data? How much detailed discussion you provide to explain the results you'd obtained (e.g., discussion about why the method works or didn't work on the data)?
3. Difficulty level - How large is the dataset used? How much effort you had to spend to collect, integrate, preprocess, and analyze the data? Are you implementing the project on a cluster or a single machine? What tools did you use (do you have to implement them or are you simply using existing libraries)?
4. Participation in the group project. How much did a team member contributes to the project.