STATS 202: Data Mining and Analysis

Instructor: Linh Tran

FINAL PROJECT

Submission due date: August 23, 2021
Report due date: August 27, 2021

Stanford University

Introduction

The goal of this class project is to give you experience in real life statistical analyses and data mining. By the end of the project, you will have learned how to identify and interpret types of different attributes in a dataset, visualize the attributes and relationships between attributes of different types, understand how those relationships could affect your models and analyses, and finally build regression models. Throughout the project, I will be available via Piazza to help out and provide advice.

You can access the *Kaggle leaderboard* to submit your predictions. Instructions for using Kaggle can be found on the *course website*. The class project is worth 200 points and is required for the course.

Background

You will be working with real world financial security *candlestick data* (at 5 second intervals) from 10 different securities downloaded from a financial brokerage firm. To prevent unauthorized downloading and use of the test data, the security, prices, and dates have all been anonymized.

The outcome being evaluted is your ability to predict the security prices up to 9 days into the future. Specifically, you will need to predict the 'open' prices for the 9 days following the end of the training data.

Data

The training data is comprised of data for 10 different stock tickers, anonymously referred to as A-J. Each of these tickers have candlestick data for approximately 87 days, provided in 5 second intervals while the stock market is open. For example, security A has a row for time 06:00:00 on day 0, which corresponds to the 5 second interval candlestick values from from 06:00:00 PT to 06:00:05 PT for the first day in the training data.

To anonymize the prices, two random values μ, σ were used to normalize the data in the following manner (rounding to the nearest cent):

$$\frac{x - \mu}{\sigma}$$

Note that, as this is real world data, there may be data quality issues related to the data such as having missing values.

Please download the data from the links provided at the *course website*. The list below describes the explicit variables included in the data:

1. *symbol* - An anonymized character indicating which of the 10 securities the prices correspond to.
2. *open* - The opening price at the begining of the interval.
3. *high* - The highest traded price during the interval.
4. *low* - The lowest traded price during the interval.
5. *close* - The closing price at the end of the interval.

6. *average* - The average traded price over the interval.

7. *time* - The time of day corresponding to the interval start, recorded in pacific time. Note that the prices correspond to the 5-second interval that follows this time.

8. *day* - The anonymized day corresponding to the candlestick prices. Note that, while financial markets typically are not open during the weekends, these values will not reflect that. Consequently, if day 5 is a Friday then day 6 will correspond to Monday. Similarly, if there is a holiday that occurs within our date range it will not be captured in these enumerated values.

Prediction submissions

Predictions should be submitted through *Kaggle*. For submissions, an identification number is given to each test set value, such that it can be identified uniquely: `‘symbol-day-time’`, where the day is enumerated starting from 0 for the first day in the test set. For example, for the first prediction for security A we have the following ID: A-0-06:00:00. Please refer to the `sample_predictions.csv` file for more examples. Note that the `sample_predictions.csv` file contains all off the intervals that you are expected to make predictions for. **Please remember to include your Kaggle team name for your submission in your write-up**. The deadline for making Kaggle submissions is Monday August 23, 2021 at 11:59 PM. Note that you are only allowed to submit up to 20 entries per day.

Evaluation

As our goal is to make future predictions, no data will be made available for the public leaderboard. Consequently, it is up to you to come up with a validation set in which you can test your models and evaluate for overfitting.

At the end of the competition, the teaching staff will be downlading the **final set of submitted predictions** and evaluating them in the following manner:

1. Predictions will be separated into two sets: days 0-3 (i.e. period 1) and days 4-9 (i.e. period 2).

2. Within each set, the squared error will be calculated for each prediction and summed across symbols. All errors will be averaged within the day, prior to being averaged across all the days in the set. Specifically, we will be evaluating in the following manner

$$\frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}}\sum_{i=1}^{10}(\hat{y}_{idt}-y_{idt})^2$$

   where $|\mathcal{D}|$ corresponds to the number of days in the set, $|\mathcal{T}|$ corresponds to the number of 5-second intervals in the trading day (i.e. 5,040), and i corresponds to the symbol number.

3. Any missing values will be omitted from evaluation.

A leaderboard will be created from the evaluated submissions and uploaded to the *course website*.

Please note that, due to the platform limitations associated with the class competition, Kaggle will attempt to display both public and private scores (which will likely be 0). We will not be relying on these evaluation scores and strongly encourage you to simply ignore them. Futhermore, while Kaggle retains all submissions made we will only be evaluating the final submission.

Write-up

As part of the final project, you are expected to submit a final report (to *Gradescope* in PDF format) covering the approaches, details, and results of the task. The report should be no longer than 10 pages (excluding figures, tables, and code) and capture the steps you took throughout the data mining process (Figure 1). Table 1 provides further details on each step of the process. **Make sure to reference your team's Kaggle leaderboard name in your report**. The code used to generate the results should be either attached as additional scripts in your final project submission, appended to the end of your report as an appendix, or referenced to in the report via a link to the uploaded git repository. Note that 10% of your grade will be based upon the organization, readability, reproducibility, and efficiency of your code.

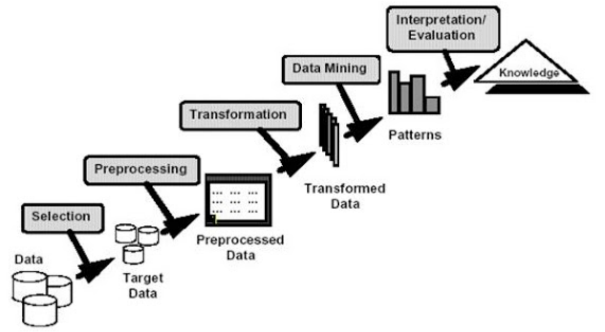

Figure 1: Steps in the data mining process.

Selection	Explain which attributes you used to build the model, and why you chose those attributes
Preprocessing	Explain whether you pre-processed any of the attributes by modifying them in any way
Transformation	Explain whether you created new features from the existing attributes, or from pairs of the existing attributes. Did you transform any of the attributes into another representation of data? Remember, you do not need to use all of the attributes in your model. Try to evaluate which attributes you think will be useful, and use those attributes.
Data Mining	Explain how you built your regression / classification model. There are many kinds of models that may work for this problem. You are welcome to use whatever regression / classification approach you would like, but remember, you need to end up with a prediction or probability.
Interpretation/Evaluation	Understand what your model is doing and how it is performing. This may require you to separate your training data into different groups so that you can test your models performance on a “hold out” group.

Table 1: Steps in the data mining process.

Your write-up should include the data processing that was done, the features used, the modeling approaches chosen, and the results of your work (**including your Kaggle team name**). Explain the decisions you made and provide visualizations supporting those decisions. Furthermore, provide visualizations in the form of tables and/or figures for each attribute in your model, and provide visualizations for pairs of attributes that you think may be related. Remember, understanding your data is an important part of the data mining process, and visualization that data can help understand it.