# COMP 4949: Assignment 1 – Juan Escalada

## Introduction

In this assignment, we will analyze a mystery dataset, and attempt to predict the value of the column “A”. The columns available, range from “A” to “Z”. It’s important to note that the data is organized in time-series format, which means each row has a corresponding Date as its index.

The best model to predict the data has been, quite surprisingly, a simple OLS Regression model that makes use of many of the columns in the data, time shifted by 1, 7, 14, 21 days. There seems to be a weekly pattern in the data, however, the Exploratory Data Analysis failed to recognize this.

## Exploratory Data Analysis

First, we will take a look at the shape of the data. Let’s see what some of the rows and columns look like:

A graph with blue lines

Description automatically generatedA screenshot of a computer

Description automatically generated

Figure 1: Data sample

The data looks a lot like stock prices, judging by the fact that their mean values are typically double and triple-digit numbers, as well as having only two decimal places (Fig. 1).

The range of the data is quite wide. The bulk of the columns have triple-digit numbers, but some reach over 5000 (Fig. 3). Most columns are very similar, so I have omitted the rest.

Figure 2: Column A over time

The values are also very volatile over time, as evidenced by this plot of our target variable (Fig.2).

A screenshot of a computer

Description automatically generatedA black screen with white numbers and a black background

Description automatically generatedThe number of columns is very large, so a correlation matrix would be unwieldy to display. Instead, I computed the correlations between all the columns, and then selected those whose correlations explain over 60% of the variance (r > 0.775).

Figure 3: Description of columns A-E

These correlations are very large and significant, showing that sometimes upwards of 85% of the variance between one of these columns can be predicted by using another column. The most highly correlated pair is A and B, with and almost perfect positive correlation (Fig. 4).

A black background with white text

Description automatically generatedI added extra columns for the values of A in previous dates, then computed the correlation. As suspected, much like stock price data, the prices of future dates are significantly correlated to the current date:

Figure 5: Correlation between A and time-shifted A

Figure 4: Correlations between Columns

A diagram of a cross between two different colored squares

Description automatically generated with medium confidenceAn easier way to visualize the correlations, is through a heatmap. Here, we can very easily see some interesting new patterns (Fig. 6):

Firstly, there is a very high positive correlation between the first 8 columns (A through H). This effect is generally not observed with most of the other columns, which actually have a negative correlation among each other, and A as well (indicated by the blue colour).

Particularly, N and O have a negative correlation with almost every other column.

If the mystery data represents stocks, this could be a direct competitor company whose stock gains when the other companies lose value.

Figure 6: Correlation heatmap

A graph with orange and purple lines

Description automatically generated

As observed in Figure 2, the data seems to spike around in a relatively seasonal fashion. Computing the moving averages for our target variable, results in an apparent trend as the months go by (Fig. 7):

A graph with blue dots and numbers

Description automatically generatedWe can see that there are around 12 spikes per year in the A column, which turns into around 6 spikes per year when the 20-day Simple Moving Average is applied, and finally into 2 spikes per year when the 50-day Simple Moving Average is computed. Although not fascinating by itself, if Column A represents the prices of a stock, this could be an excellent indicator of its seasonality, and therefore a great basis for a predictive model that leverages time-series data.

Figure 7: 20 and 50-day SMA

Another technique for identifying trends, is autocorrelation plots. Here, we will compare the correlation of the current price to its past prices (Fig. 8). Once again, there are some interesting findings: PACF shows a significant positive correlation between the current value of A, its value 1 day ago, 8 days ago and 22 days ago. These are prime candidates for a predictive model.

ACF, which does not adjust for the time differences between correlations, is good for seeing whether an ARIMA (Auto Regressive Integrated Moving Average) model would be relevant, however, it does not help us decide which features are the best predictors.

Figure 8: Partial autocorrelation plot for A

Based on these observations, the most relevant variables are all the columns from B to H, as well as the time-shifted A columns from 1, 8 and 22 days ago. These are likely to be the most helpful in creating a strong model.

## Model Development

Based on the exploratory data analysis, we found various relevant variables to build our models. We will be building three different models to see which one best predicts the target variable.

* Model 1: A simple model that uses Linear Regression to predict the next values of A. It makes use of KNNImputer to fill in the missing values. The model makes use of the values of A from 1-, 7-, 8- and 14-days prior, as well as various other time-shifted columns, typically with a 1-, 7-, 8-, 14-, 21- or 22-day lag. The reason why we do this, is because we cannot use the current values for those columns (which would not normally be available when making future predictions).
* Model 2: This is another relatively simple model that makes use of Holt-Winters Triple Exponential Smoothing Forecast to find seasonal patterns and apply that to our predictions. In this case, we only use the value of A (in the past) as a predictor for the future values.
* Model 3: A simple ARIMA (Auto Regressive Integrated Moving Average) model which makes use of time-shifting on A for predictions. We used 1-, 7-, 14- and 21-day time-shifted values of A as the predictors, which were the ones with the highest predictive power.

## Model Evaluation

All our models performed rather mediocrely, as none of them managed to get under the 300-mark for the RMSE. Initially, I was getting values of under 100, but this is because I was not using time-shifted columns for my predictions (which doesn’t make sense).

The evaluation is the Root Mean Squared Error (RMSE) among the test data, which is the last 10 dates provided. Values of around 300, are not terribly inaccurate, but not too trustworthy when the vicinity has lower values. Remember that the data has occasional spikes which means that a large variance will be less noticeable when the vicinity has large values as well.

The best model, rather ironically, has been the OLS regression model. We have also attempted to use a deep learning model for predictions, however, once again, we got very low RMSEs (under 100), which may be an indication of overfitting on the test data. To err on the side of caution, I picked the OLS regression model to be more likely to predict accurately outside of the test data.

The following chart summarizes the models’ performance.

|  |  |  |
| --- | --- | --- |
| **Model Type** | **Features** | **10-day RMSE** |
| OLS Regression | A\_lag\_1, A\_lag\_7, A\_lag\_8, A\_lag\_14, C\_lag\_21, D\_lag\_21, E\_lag\_7, F\_lag\_21, H\_lag\_7, I\_lag\_21, J\_lag\_7, L\_lag\_14, L\_lag\_21, M\_lag\_21, M\_lag\_22,  P\_lag\_22, Q\_lag\_2, Q\_lag\_21, S\_lag\_2, S\_lag\_7, U\_lag\_7, U\_lag\_22, W\_lag\_21, Z\_lag\_2 | 308.256 |
| Holt-Winters Triple Exponential Forecast | A | 327.413 |
| ARIMA | A\_lag\_1, A\_lag\_7, A\_lag\_8, A\_lag\_14, A\_lag\_21 | 321.036 |

The following chart shows the Predicted values of A versus the actual values by using the OLS Regression model (Fig. 9). We notice that it consistently predicts a bit higher than the actual values. Despite my best efforts tweaking it, it did not yield better results by removing some of the variables with large positive effect on the final predicted value. Furthermore, tweaking it to fit the test data would only result in overfitting which would just decrease the predictive power even if RMSE were lower.

A graph with numbers and lines

Description automatically generated

Figure 9: Predicted values of A versus actual values for OLS Regression Model

## Summary

In summary, the OLS regression model was the best according to the RMSE metric. That being said, it can be very hard to come up with a truly accurate model, as each has its own strengths and weaknesses.

There are many relevant features, almost all of which have a p-value of less than 0.001. This means that a machine learning model capable of integrating all these parameters, could potentially have excellent performance. The problem is, that time-series data is harder to model with deep learning models as the training data must be relatively recent (or otherwise back-testing must be performed).