DATA ACQUISITION AND PROCESSING SYSTEMS ELEC0136 22/23 REPORT

SN: 19076187

ABSTRACT

This report discusses the process of collecting auxiliary data for the purpose of creating a stock forecasting model for the American Airlines (AAL) stock, the data exploration, preprocessing and inference stages. From this a model is created which was found to not predict the stock price effectively, however was able to capture the trends that the auxiliary data had on the stock price, however to an incorrect scale with a root mean squared error of 1.385.

1. INTRODUCTION

This project is focused on analysing the market trends of American Airlines Group Inc (AAL) and providing advice on when and whether to buy, hold, or sell the stock. The project aims to use a machine learning data-driven approach to understand the performance of American Airlines and the impact of broader economic conditions on the stock, the impacts of the pandemic and commodity prices. We will be using various types of auxiliary data to achieve this goal such as jet fuel prices, Consumer Price Index (CPI) and COVID data.

These auxiliary data and the primary historical AAL ticker data are used to train a Long Short-Term Memory (LSTM) model for the purpose of forecasting the AAL stock price. Through the results and evaluation of this model a decision will be arrived upon weather to buy, hold or sell the stock.

This report will be split into 9 further sections which include: **Data Description** which will detail the primary and auxiliary data has been used as well as the reasoning behind it. **Data Acquisition** which will detail the methodology behind acquiring the primary and auxiliary data. **Data Storage** which will explain the data storage method and the reason it was chosen. **Initial Data Preprocessing** which will detail the initial processing necessary for data exploration**. Data Exploration** which will detail how the data was explored to gain a better understanding of the dataset to inform the data preprocessing stage**. Final Data Preprocessing**, which will detail the final preprocessing steps such as data outlier cleaning and data transformation used to make the data suitable for analysis which have been justified within the data exploration stage. **Data Inference**, which will detail and justify the model selection, training and evaluation. **Conclusion** which will give an overview of the results, stock decision and future considerations for the model and analysis. **References** which will be used to provide context to claims made.

2. DATA DESCRIPTION

This section of the report will detail the primary and auxiliary data used in this report through description of content, size and format of the data as well as the reason for selecting the auxiliary data. The primary data is the historical AAL stock prices and the auxiliary data is jet fuel prices, COVID data and CPI data from the start of April 2017 to the end of April 2022.

2.1. AAL Stock Data

Historical AAL stock data is the primary data as it contains the variable we aim to predict from our forecasting model which is the opening stock price. The AAL stock data will be daily from the start of April 2017 to the end of April 2022 and will contain the daily opening, high, low and closing prices as well as the daily trading volume. The opening and closing price provide information on the market sentiment at the beginning and end of the day which can be used to predict the direction of the stocks movement. The high and low prices represent the highest and the lowest prices which the stock traded at during that day, which indicate the volatility of the stock which may be used to predict how much the price may fluctuate in the future. Trading volume represents the number of shares of a stock that are traded during the day which provide information on the level of interest in the stock which may be used to predict future demand. Using historical data from 01/04/17 to 30/04/22 will provide 6400 observations which will be in a float68 format which will be approximately 90KB when stored in CSV format.

2.2. COVID data

COVID data was selected to be used as an auxiliary data over this time period as it played a significant impact on the airline industry as well as the economy as a whole, which will be able to provide the model with context regarding change in the historical AAL stock data over this period. Specifically, domestic COVID total deaths and risk level data are used for the forecasting model as they able to understand how the pandemic affected demand for domestic air travel, which is the primary source revenue for AAL [1]. This data is

important as travel restrictions and lockdowns, which are often based on the total deaths and risk level, have a direct impact on the demand for domestic air travel. It must be noted that this data only exists from 09/03/20 as this is the date when American was first affected by COVID, therefore this must be taken into consideration in the preprocessing stage. This dataset contains two variables, total deaths which is the cumulative number of deaths across American confirmed to be by COVID, and the risk level which is a measure of the risk of COVID. This risk is calculated into 4 discrete values ranging from 0 to 3 which correspond to: minimal, low, moderate and substantial risk which take into account the number of number of new cases per 100,000, COVID test positivity rate and number of hospitalisations per 100,000. The format of these two variables is float68 and within the given time period contains 1568 none-zero observations which correspond to a file size of 17KB stored in a CSV format.

2.3. Consumer Price Index (CPI) data

The Consumer Price Index (CPI) and Consumer Price Index for Urban Wage Earners and Clerical Workers (CPI-W) have also been used as auxiliary data. CPI are measures of the changes in prices of goods and services that households purchase, which is used to measure the average change over time of prices paid by consumers for a basket of goods and services. CPI-W is specifically for goods and services that a urban wage and clerical workers for a specific set of goods and services. The relevance of CPI and CPI-W to the stock price of AAL is that these measures are able to provide insight into inflation and economic condition, which will be direct impacts on AAL stock. Inflation can increase the costs and services that airlines reply on, such as labour, which in turn affect the financial performance of the company which has direct effects on the stock price. Also economic conditions can also impact the demand for air travel, as consumers may be more likely to travel during times of economic prosperity and less likely to travel during times of economic downturn. It must be noted that CPI and CPI-W are monthly measurement, therefore this must be taken into account in the preprocessing stage. This data will be monthly and will contain a CPI and CPI-W value which are both of type float68. This data will contain 148 observations for the given time period and will take 2KB of storage in CSV format.

2.3. Kerosene Jet Fuel Prices for U.S. Gulf Coast data

Kerosene jet fuel prices have also been used as auxiliary data as jet fuel is a significant cost for airlines roughly making up 20-25% percent of total costs, therefore the change in fuel price will have a direct impact on the profits of AAL and subsequently the fuel [2] **.**Specifically fuel prices from the U.S. Gulf Coast are used as the domestic flights are the primary revenue source for AAL. Fuel price auxiliary data was selected as it has a direct impact on AAL's profits which is a key statistic which investors look at. This data is daily over the given time period and will contain 1326 observations, with a csv file size of 22KB.

3. DATA ACQUISITION

The data acquisition method of all of the primary and auxiliary datasets was done using APIs, primarily due to their ease of use, efficiency in terms of data preprocessing, and flexibility in accessing large amounts of data from a variety of sources.

APIs provide a consist format for the data and automate the data collection process, which eliminates the need to manually download or to write web scraping scripts which require extensive preprocessing. They also allows for more flexibility in data acquisition, as you are able to access multiple datasets from multiple sources and different types of data using the same API.

API's, or application programming interfaces, are a method of retrieving data from web-based services or databases through a request to a specific URL endpoint. APIs allow for the automation of data collection, eliminating the need to manually download or complex web scraping methods which require significantly more set up. Additionally in terms of data quality APIs provide a cleaner and more consistent structured data than web scraping. Web scraping can be prone to errors and inconsistencies as it involves extracting data from websites without the explicit consent of the website owner. Significantly, web scraping can raise ethical concerns such as privacy, copyright infringement, etc whereas APIs do not as they require authentication and authorization to access the data which the data provider is aware of.

In summary, web APIs were exclusively used for the primary and auxiliary data collection as these APIs were readily available with the pre-setup infrastructure and documentation for each auxiliary data desired, making the implementation of web scraping unnecessary.

3.1. AAL Stock Data

The AAL stock was acquired using the yfiance library. The yfiance library was used as it required minimal code, no API key and offered daily stock data for free. It was used to acquire daily AAL stock data containing the opening/closing prices, daily high/lows and the trading volume.

3.2. Covid Data

The covid data was collected using the Covid Act Now API, which required an API key which could be obtained through creating an account and detailing the use of the API. This API was used due to its extensive documentation explaining the

data as well as endpoints as well as allowing data from specific regions. This API also allowed for CSV rather than JSON meaning less preprocessing was required. This API was used to obtain total deaths and risk level across America.

3.3. Consumer Price Index (CPI) data

The CPI data was obtained through the Bureau of Labour Statistics (BLS) and required an API key which was obtained through account creation. This API was used as it is the primary government source for CPI data and had extensive documentation and variations on CPI such as CPI-W. Note the CPI data is only available in JSON, therefore it must be processed into CSV format which is used.

3.4. Kerosene Jet Fuel Prices for U.S. Gulf Coast data

The Federal Reserve Economic (FRED) was used for the jet fuel data as it offered daily historical data in a structure format. The FRED API required an API key which was obtained through account creation. This API was used primarily due to its ability to provide daily historical data compared to manual downloads.

4. DATA STORAGE

The data storage method I used was CSV. This was primarily chosen due to easy of human readability, ease of manipulation through compatible libraries, and being able to store the files locally.

CSV or 'Comma Separated Values' is a file format which allows for data to be stored in a tabular form, where each row represents a record and each column represents a field of the record. Due to this simple design CSVs are easy for humans to read, allowing for quick sanity checks upon changing the data, looking for missing values or noticing incorrect data. Another benefit of the CSV format is its ease of manipulation due to it's compatibility with multiple technologies, allowing for the flexibility of using multiple tools without having to convert to a different intermediate format. This is useful when you need to perform data cleaning, transformation, or analysis on the data.

Additionally, I used CSV format because it is safe to store files locally. This is because CSV files are stored in plain text format, which means that they can be easily backed up and protected against data loss. As the data is small (max 100KB total), it is also easy to keep a record of the files and not lose them.

On the other hand, other formats such as pkl, numpy or cloud based storage services have their own benefits, but they may not be suitable for our specific case. The pkl format is useful when you need to store large amounts of data in a compact format, but since our data is small, it is not required. Numpy format is useful when you need to perform mathematical operations on the data, but this can also be achieved through converting csv to dataframes on pandas. Cloud based storage services are useful when you need to store large amounts of data and share it with others, but since our data is small and is not worked on in a celebrative way this is not required. Cloud based storage also require additional code implementation which again would be unnecessary in our use case.

5. INITIAL DATA PREPROCESSING

This section will detail the initial data preprocessing which is required to put the data into the same format which can then be explored and visualised. As there are several different auxiliary datasets being used and they are acquired in slightly different formats. This section will detail the preprocessing applied to standardise the data into 1 cohesive dataset to be analysed. It must be noted that as the data comes for API's the data may have already been cleaned and standardised to some extent, especially the BLS API as it is a primary source for this data. Note that all the axis some axis are arbitrary and graph are used to visualise trends and patterns.

Firstly, it was noted in the data acquisition stage that CPI and CPI-W data is monthly with a year column and a period column. To standardise the date index on this csv firstly we convert the year and the period column into one date column on the first of each month. Then we expand the dates which are only for the first of each month from 2017-2022 to daily by copying the CPI and CPI-W to be the same the entire month. This assumes that the CPI as the CPI value does not seem to jumpy massively month by month as can be seen by figure 1.

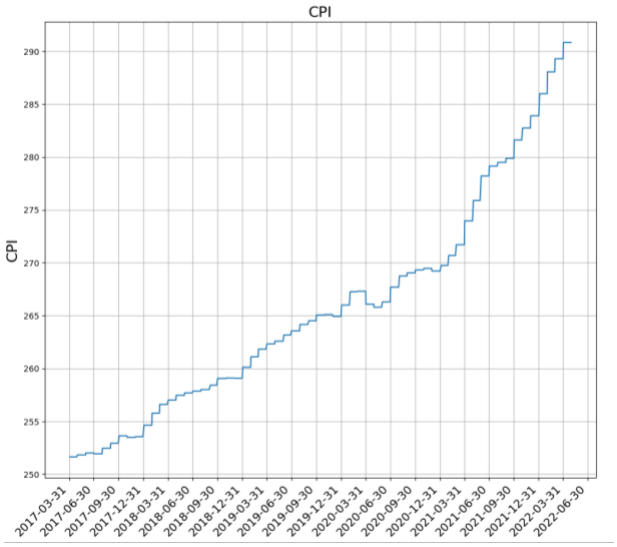

Fig. 1 CPI against Time

The AAL stock data date format must also be fixed as it contain the time as well as the date in the 'Date' column. To fix this string splicing is used to remove the unnecessary time from each row in the 'Date' column.

The COVID data is also expanded as the COVID data starts on the 2020/03/09 with 0 deaths and 0 risk we can extrapolate this back to 2017/04/01 as we know there were no COVID deaths before 2020/03/09 and therefore the risk levels were also 0 as shown by figure 2.

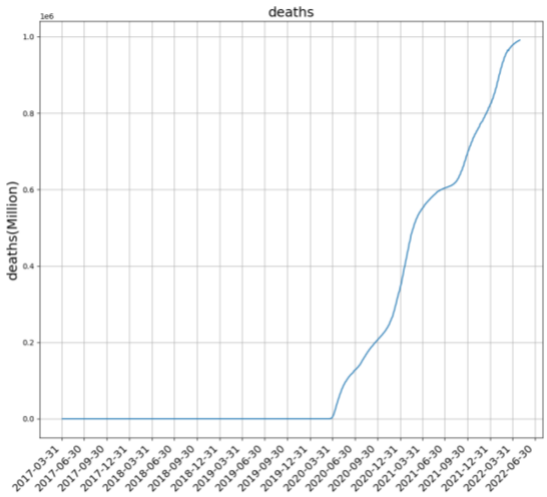

Fig. 2 Total COVID deaths (millions) against Time

We can also note some missing data point values in the values found in the jet fuel price dataset. To ensure that the dataset is consistent throughout we will interpolate the missing values. We are able to make this assumption as the data is missing at random, the underlying relationship between the variables is relatively smooth and consistent, the distribution of the data is relatively stable and the missing values are not indicative of any underlying trend or pattern:

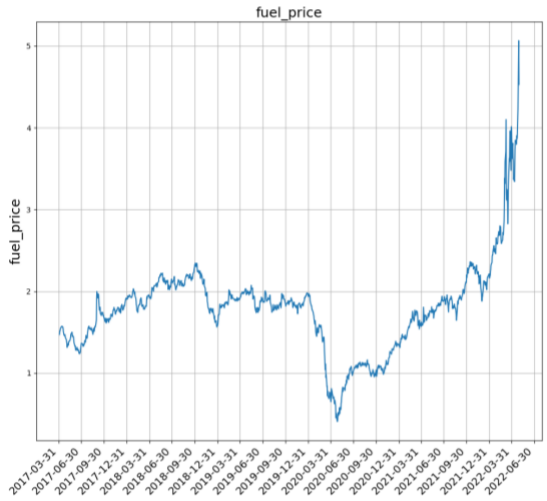

Fig. 3 Kerosene Jet Fuel Prices against Time

This graph shows a relatively linear smooth graph other than two possible outliers from 2019-12 to 2020-09 and 2022-03 to 2020-06. These outliers will be explored further in the data preprocessing stage.

6. DATA EXPLORATION

Following the data combination into a single csv where the data has been cleaned and formatted, this section will detail the data exploration methods and the patterns and conclusions drawn. It must be noted that data exploration has been put before the final preprocessing section, as the final preprocessing step uses Principle Component Analysis (PCA) which will transform the data leading to patterns being obscured and relationships skewed. Additionally PCA changes the variables to principle components which are harder to gain insights from. It must be noted that the data exploration and preprocessing stages are iterative therefore, it is difficult to describe in individual sections, hence the following structure.

5.1. Exploratory Data Analysis EDA

Firstly, we will test out data for normality as it is a condition for using z-score to finding outliers, as the z-score should only be used on data which follows a normal distribution otherwise there will be misidentification of outliers. For this purpose a function called test_normality() which takes in the file path of the combined.csv and tests each column (which correspond to each variable in each dataset) for normality using three tests and returning a table listing if the test found the column to be normally distributed or not at a 5% significance level. The following tests are used:

In Shapiro-Wilk test the null hypothesis is that the sample is drawn from a normal distribution. The test statistic is calculated from the sample data and compared to the critical value from the table of the Shapiro-Wilk distribution. If the test statistic is greater than the critical value, the null hypothesis is not rejected, and the data is considered to be approximately normally distributed.

The Anderson-Darling compares the sample data to the normal distribution. The test statistic is calculated from the sample data and compared to the critical values from the table of the Anderson-Darling distribution. If the test statistic is greater than the critical value, the null hypothesis is rejected and the data is considered not to be normally distributed. It must be noted that the Shapiro-Wilk test is known to be sensitive to deviations from normality in the tails of the distribution. It should be noted that the Anderson-Darling test is known to be more sensitive to deviations in the middle of the distribution.

Lilliefors is a variation of the Kolmogorov-Smirnov test for normality specifically for small sample sizes. This test uses a statistic based on the Kolmogorov-Smirnov statistic,

which compares the sample data to the normal distribution. The test statistic is calculated from the sample data and compared to the critical values from the table of Lilliefors distribution. If the test statistic is greater than the critical value, the null hypothesis is rejected and the data is considered not to be normally distributed. It should be noted that the Lilliefors test is sensitive to deviations in the centre of the distribution.

By using multiple tests, the chances of detecting any deviations from normality increases. Here are the results from running this function:

Test	anderson	lilliefors	shapiro
Column			
CPI	Not Normal	Not Normal	Not Normal
CPI_W	Not Normal	Not Normal	Not Normal
Close	Not Normal	Not Normal	Not Normal
High	Not Normal	Not Normal	Not Normal
Low	Not Normal	Not Normal	Not Normal
Open	Not Normal	Not Normal	Not Normal
Volume	Not Normal	Not Normal	Not Normal
deaths	Not Normal	Not Normal	Not Normal
fuel_price	Not Normal	Not Normal	Not Normal
risk_level	Not Normal	Not Normal	Not Normal

Fig. 4 Table of all variables normality test results

From these results it can be seen that none of the variables follow a normal distribution, therefore the z-score is not a applicable method to test for outliers, therefore interquartile range will be used.

Correlation is a statistical measure which describes the relationship between two variables. Typically, in the data exploration stage variables are tested for correlation as it useful to identify patterns and relationships within the date. The correlation of each variable was tested against each other to produce a heatmap to provide an overview of which variables impact each other:

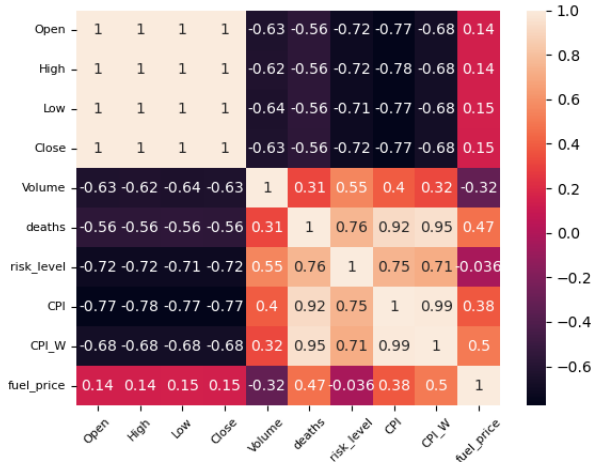

Fig. 5 Correlation matrix of all variables

There is an expected 1 correlation between open, high, low and, close as they directly impact one another. This graph also notably shows that the stock price is negatively correlated with the number of deaths and risk level during COVID which is expected as COVID progressed these values increased and the economy as well as travel fell. It is notable that CPI is more strongly correlated with the stock price than CPI-W, which may imply that wage earners and clerical workers are less influential the stock than the general population. This is likely as CPI-W is a more specific indicator of inflation. It is also notable that fuel price is slightly correlated with the stock price which may be due to the general increase in fuel price overtime and the general stock price increase overtime. This chart also lets us see the correlation within the auxiliary data showing that risk level COVID was weakly/not correlated with fuel price suggesting that COVID did not influence the kerosene market. It must also be noted that these correlations reinforce the linearity of the data such can be seen when the data is visualised into scatter graphs.

5.2. Hypothesis testing

Hypothesis testing is a statistical method used to make inferences about a population based on a sample of data. It is often used during the data exploration stage to test for differences or similarities between groups, or to test for a specific relationship between variables. A t test is a type of hypothesis test that is used to compare the means of two groups and determine if they are dependant or independent. We will use t tests to test if the mean stock price changes in periods of high and low auxiliary data at a 5% significance level.

For our first test we will t test to see periods of high CPI or low CPI has a significant impact on the means of the AAL stock opening price within high and low periods. We divide the CPI into high and low periods using periods which are above and below the mean of the entire group.

The output from the CPI t test: t-statistic: -46.59 and p value: 1.09e-277. The t statistic is a measure of the difference in the means of the stock price using standard deviations. Therefore, as the means are 46.59 standard deviations part and the p value is so low we can say which means that there is very strong evidence to suggest that there is a significant difference in the mean AAL stock price during periods of high and low CPI inflation. Furthermore, the t-statistic is negative, which means that the mean AAL stock price is lower during periods of high CPI inflation than during periods of low CPI inflation.

We also performed this t test on the number of COVID deaths following the same method to find t-statistic: - 24.58 p-value: 1.42e-109 . From these statics we are able to see that as the t statics is largely negative and the p value is also very small that there is a significant difference between the

mean stock price during high and low COVID deaths, specifically that the average AAL stock price was lower during high periods of COVID deaths. This is likely due to COVID deaths being correlated with the economic climate and travel restrictions.

The same test was also performed on fuel price showing: t-statistic: 6.26 and p-value: 5.39e-10. Which leads to the same conclusion periods of high clue prices lead to lower mean stock prices. This is likely due to increased costs and decreased profits.

7. FINAL DATA PREPROCESSING

In this section I will summarize the final data processing before the data is input into the forecasting model. This section will detail the process of finding and dealing outliers and the data transformation method and the reasoning behind it.

Firstly, as established in the data exploration stage none of the data follows a normal distribution, meaning it is not accurate the use z score to search for outliers. Therefore, outliers were classified through calculating the first and third quartiles and the interquartile range and calculating the upper and lower bounds for outliers as 1.5 times the IQR away from the first and third quartiles. This method flagged CPI, volume and fuel prices. Further research must be done to determine whether it is appropriate to remove these outliers or not depending on their cause.

The CPI was flagged for outliers from October 2021 till April 2022. Upon further research this is due to economic regulation due to the impact of COVID 19. This is due to the fact COVID 19 had a significant impact on the global economy leading to government intervention such as stimulus packages, interest rate cuts and changes to public trade policies. These actions can have a significant impact on inflation and consumer prices. Additionally, disruptions to supply chains were caused during the pandemic leading to shortages of certain goods and services leading to prices rising. Due to these reasons we observe the outliers in the CPI data. As the increase in CPI has legitimate cause I have decided not to remove the CPI outliers and this data is able to provide greater insight into the broader context of the market on the performance of the AAL stock market .

Volume from 2020-03-24 to 2021-03-15 was also classed as an outlier. There are several possible reasons for this spike in trading volume. Firstly, due to the increase in the number of retail investors [3] due to stock trading becoming more popular and accessible primarily due to the introduction of phone applications such as Robinhood. Additionally, during COVID there was rampant speculation on the markets which combined with quarantine and the increase in retail investors lead to significantly increased trading volume. For these reasons I plan to keep the increased volume within the dataset as the trading volume can be heightened during moments of heightened speculation and it would be useful to train our model with this information.

Fuel price from decrease from 2020-03-18 till 2020-05-04 and a rapid increase from 2022-02-11 till 2022-04-29. Firstly, the outlier decrease from in 2020 was likely due to the decrease in fuel demand throughout COVID which lead to a decrease in fuel prices. Additionally, there was also a price war between Saudi Arabia and Russia over this period which lead to further declines [4]. The increase in fuel price in 2022 was reactionary to the pandemic where many refineries closed leading to a decrease in volume of sales leading to an increase in the cost. This was also further exacerbated by the start of conflicts in eastern Europe with Russia [5] as Russia is a key exporter. I have decided to keep both these outliers in the data as keeping these outliers provides important context to the performance of the AAL stock during this period. Additionally, this information provides a greater understanding of how the stock price reacts or adapts to major fuel price shifts.

5.2. Data Transformation

Data transformation is the process of manipulating and alter the structure, format and values of a dataset for the purpose of making it for suitable for a particular analysis or modelling technique. These techniques include operations such as normalization, scaling and dimensionality reduction. The goal of data transformation is to improve the performance and accuracy of a model through making the data more consistent, informative, and robust.

In the data transformation stage we applied the following changes: normalisation, scaling and Principle Component Analysis. The primary transformation was PCA which is used to transform the data into a lower dimensional representation. PCA helps identify patterns and relationships which are not immediately obvious, through the reduction of dimensionality PCA allows us to make relationships between different variables clear and interpretable. This is particularly useful for time series data. The prerequisites for PCA are correlation, normality, linearity and scale. In order to use PCA we must first check if the dataset is correlated, and which was done during the exploration stage and found the data to be correlated. This is because PCA can only applied to correlated data as this is used to identify the principle components. If the data was not correlated techniques such as random project or Linear Discriminant Analysis (LDA) could be used. Secondly, as noted in the exploration stage the data is not normally distributed, therefore min max normalisation is applied as PCA assumes normality. During the exploration stage we are able to see that all of the time series graphs follow a rough linear

relationship. Finally, we must scale the data as we must ensure all data is on the same scale before applying PCA.

In conclusion, the use of PCA is used for our forecasting model for AAL as it ensures robustness and accuracy through the reduction of dimensionality and redundancy of data allowing for a better fitting model which will be better suited to avoid overfitting.

8. DATA INFERENCE

In this section, we will address the problem of predicting the opening stock price of AAL using historical auxiliary data. The model selected for this task was a Long Short-Term memory (LSTM) model. LSTMs are a type of recurrent neural network such are well suited for time series data due to their ability to make used of past information for perditions which is inherent in stock time series data. PCA was particularly useful for the LSTM model as it helps improve the performance and reduce of overfitting as LSTMs are prone to overfitting with high dimensional data as they are sensitive to a large number of input features. An LSTM model is also able to handle multiple inputs variables unlike linear regression, decision trees or k nearest neighbour models.

The model was implemented using the keras library and was built with multiple layers including an input layer, LSTM layer and output layer. The number of layers was based on iterative process to find the best combination which lead performance on the test set. The final model was trained using the preprocessed data and used the mean squared error (MSE) loss function which penalises large errors more heavily which is well suited for our forecasting problem. The Adam optimizer was chosen as it is computationally efficient and has been shown well to work in practise. The Adam optimizer uses an adaptive learning rate which is especially useful for helping the model converge faster. The dataset is split into a 70 30 train test split.

The evaluation was based on root mean squared error (RMSE), mean absolute error (MAE) and the r-squared score (R2). Firstly, RMSE is a measure of the difference between predicted and true values using the square root of the mean of the squared differenced between the predicted and true values, where a lower RMSE indicates a better fit. Secondly, MAE uses the absolute differences between predicted and true values, where a lower values indicates better fit. R2 is a measure of how well the model fits the data, ranging from 0 to 1. A higher values indicates a better fit and 0 indicates a model no better than a model which predicts the mean of the target value. These metrics where chosen as they are better suited for predicting continuous values as compared to metrics such as accuracy which does not take into account the magnitude of the errors, simply a binary correct or not correct.

Here are the results of our model:

	Metrics		
Results	RMSE	MAE	R2
Training	0.049	0.027	0.965
Test	1.384	1.225	-19.851

Fig. 6 Table of model results on training and test set

Firstly, these results show that the model did not generalise well as suggests overfitting due to the training metric scores being significantly better than the test metrics. We note a significant difference with each metric compared to its test counterpart. The difference in RMSE and MAE suggests that the test model was a significantly bad fit as a lower value indicates a better fit. Secondly, we note that the R2 value is significantly negative, suggesting that the model performed worse than a model which predicts the mean of the target value.

Another evaluation method for the performance of the model was plotting the prediction labels against the actual labels. This allowed the visual comparison of the predictions against the actual test set to get a sense of if the model was able to predict any trends of the data for the test set.

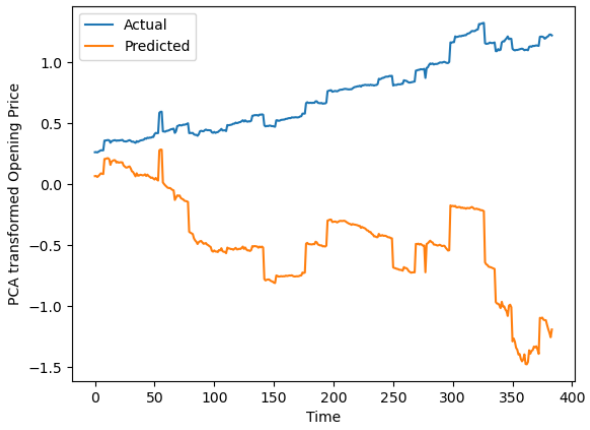

Fig. 7 Graph of actual stock opening price vs predicted

This graph shows that the model was infect able to roughly capture some of the trends within the actual labels, albeit in the wrong direction. We note the spike captured at time 50 and stock being stable before a decrease at 140 as some of the more apparent trends being captured. It seems as though the model has been able to predict the trends and underlying patterns in the stock price however in an incorrect scale.

This is likely done to the model not correctly learning the model not correctly learning the extent to which the auxiliary data affect the data, however shows that the model was able to learn the overall affect it had on the stock price.

Overall, out model did not perform well as performed worse than a model which just predicts the mean of the target values. We must however note that this model was able to capture the trends and underlying patterns that the auxiliary data had on the AAL stock price, however not to the correct scale. This is due to the model not generalising well and being over trained. These issues may be overcome with more training data over a longer period of time and to a higher precision, more auxiliary data types which capture a wider array of factors such as political events, public sentiment and more economic conditions. It must be noted that stock price forecasting is an inheritably complex problem to predict to a high degree of accuracy, therefore these results still performed better than expected as they seemed to capture the trends of the stock price time series.

9. CONCLUSION

In conclusion, the model was unable to accuracy predict the stock price using the auxiliary data of COVID, CPI and, jet fuel prices which was expected however the stock was able to roughly capture trends in the stock price at an incorrect scale. In terms of the main goal of this project to decide whether the stock of ALL should be sold, held or bought the model does not give a conclusive result for this and we must turn to our exploratory data analysis. The analysis showed a negative correlation with COVID risk level and deaths, CPI and fuel prices. Therefore, as we know that we are nearing recovery from the pandemic and therefore CPI is set to recover all we can assume the stock price will gradually increase overtime, therefore I would put the stock as a ‘buy’. This suggestion does not take into account other economic factors such as further economic metrics, political events and sentiment. It must also be noted we are using historical data and the process of predicting this auxiliary data may be as difficult as predicting the stock price itself. Therefore this model is not able to predict the future stock price and is mainly used to explore the patterns and trends between the historical auxiliary data and stock price.

For future considerations we may look into getting higher precision data such as hourly AAL stock data, however these are locked behind pay walls, public sentiment through web scraping, however this must be done in an ethical way and more economic indicators such as GDP and employment rates.

10. REFERENCES

[1] “American Airlines Group's passenger revenue in FY 2021, by region,” 12 12 2022. [Online]. Available: https://www.statista.com/statistics/422399/operating-revenue-by-region-of-american-airlines-group/.

[2] R. luman, “Soaring fuel prices complicate aviation sector’s recovery from the pandemic,” THINK Economic and Financial Analysis , 21 02 2021. [Online]. Available: https://think.ing.com/articles/soaring-fuel-prices-complicate-aviation-sectors-recovery-from-pandemic. [Accessed 12 12 2022].

[3] B. Pisani, “Trading volume is up from 2020′s breakneck pace as retail investors jump in,” CNBC, 22 01 2021. [Online]. Available: https://www.cnbc.com/2021/01/22/trading-volume-is-up-so-far-from-2020s-breakneck-pace-as-retail-investors-get-even-more-active.html. [Accessed 10 01 2023].

[4] K. M. Camp, “From the barrel to the pump: the impact of the COVID-19 pandemic on prices for petroleum products,” BLS, 10 2020. [Online]. Available: https://www.bls.gov/opub/mlr/2020/article/from-the-barrel-to-the-pump.htm. [Accessed 12 01 2023].

[5] R. Pipoli, “ Strong dollar, visa delays help keep 2022 jet fuel volume still below pre-pandemic levels,” Reuters, 25 10 2022. [Online]. Available: https://www.reutersevents.com/downstream/workforce-development/strong-dollar-visa-delays-help-keep-2022-jet-fuel-volume-still-below-pre. [Accessed 13 01 2023].