Analyzing Trends Between Crude Oil Prices and Airline Statistics

Aditya Kalyan Jayanti
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586
aj8582@cs.rit.edu

***Abstract*— The airline industry is a competitive market and has a small profit margin. In the past year, twenty-four airlines worldwide have ceased operations, and major problems reported by them are high oil prices, market fluctuation, financial troubles, and poor customer service. Furthermore, flight delays, cancellations, and over-bookings impacts cost to the customer and airline companies. The purpose of this analysis is to develop an early warning system by analyzing airline performance with crude oil prices using forecasting algorithms and natural language processing (NLP). Analyzing and accurately forecasting time series data is a challenging task due to seasonal variation and non-linear trends. Experiments conducted in this paper also describe the relationship connecting the sentiment of news articles and crude oil prices. This paper discusses approaches to analyze major factors affecting airline profitability in three parts: 1) trend analysis of crude oil prices using internal and external factors; 2) predicting flight delays and computing operating costs as a result of delays; 3) analyzing the effect of high operating costs and oil prices on airlines.**

***Index Terms*—Natural Language Processing; Crude Oil Prices; Machine Learning; Airline Performance; Time Series Analysis.**

I. Introduction

Crude oil is an energy source that is extremely important to the world's energy markets and has a significant influence on macroeconomics. It affects industries including [27] transportation, manufacturing, and heating. Any alteration in production costs directly impacts consumers. Since jet fuel is a major petroleum product produced by processing crude oil, these prices are often correlated. The biggest plunge in oil prices after the 2008 financial crises occurred during mid-2014 [3] this sharp drop resulted in increased profits for airlines, which led to revamping their entire operating fleet [32]. Fuel costs constitute about a third of airline operating costs and hence, have a significant impact on profitability. In a recent analysis by MIT engineers and Wall Street analysts on the profitability of American Airlines (AA) determined that airline fuel costs are extremely sensitive to every dollar change in crude oil prices [32]. An interactive dashboard developed by [32] provides a more insightful view in determining the sensitivity of airlines and crude oil prices. Between 2016 and 2018, oil prices increased by $21 per barrel, this increased AA fuel costs by $2674 billion.

For airlines, the price of oil is the most significant expense after payroll, and due to this, some airlines often involve in hedging fuel prices [12] to shield them from fluctuating costs. On the contrary, as demand for oil collapses in 2020 along with most airlines reducing capacity by 90% due to ongoing pandemic, it results in the opposite effect. In some cases, airlines lose up to $1 billion dollars [14], attributed to fuel hedging loss. Flight delays and cancellations are other factors that cost airline carriers billions of dollars each year. In 2007, air transportation delays in the US cost airline companies $8.3 billion and reduced the US GDP by $4 billion [4]. The above factors affect the airline industry and the US economy greatly. Hence, forecasting trends of crude oil and analyzing flight delays provide insights to sustain businesses through extreme volatility especially during an unprecedented time.

The data required for part 1 of this project is ordered as a function of time hence, we perform time series analysis using forecasting tools. In machine learning, data is usually split between training and testing with a ratio of 70:30 or 80:20. But, because of the nature of a time series (e.g., seasonality, non-linear trends), it cannot be separated into distinct ratios randomly [11]. For this reason, training and testing data must include non-overlapping periods. In part 2 of this project, using airline on-time performance data, the flight delays are analyzed and compared with crude oil price trends. Finally, this paper contrasts between historical and current events involving the profitability of the airline industry.

The report has been organized as follows. section II provides a brief background on the factors affecting crude oil prices and their relation to airline operating costs. Section III describes the data used in this project in three steps. Section IV provides a description followed by various tools used in completing this project. Historical crude oil is analyzed in section IV-A. Followed by analyzing sentiment of news articles related to crude oil in section IV-B. Section IV-C provides an analysis describing the impact of oil on airline profitability. A forecasting tool to analyze crude oil trends is described in section IV-D. The accuracy of this forecast using cross-validation is presented in section IV-E and the error rate is visualized in section IV-F. Exploratory data analysis is performed in section IV-G. In section V, we describe the results and their limitations. Conclusions and future work have been elaborated in section VI.

II. Background

Several computational techniques used in prediction algorithms include neural networks, support vector machines (SVM), and time series forecasting. But, the popularity and wealth of information available through social media sites like Twitter and news stories from the mainstream media have prompted several researchers to use online media data to build forecasting models. An article published in the International Economic Review [9] describes an early stage proof of concept on how stock price movement relates to media information.

One such implementation described in [19] determines that considering only the title and ignoring the content of the news articles would result in reduced prediction performance. But, unlike the title, the content of a news article may contain several commonly occurring words that require preprocessing while performing NLP tasks. Processing hundreds and thousands of news articles increase the time complexity and would require additional computational resources. However, ignoring the content of an article increases the chances of missing valuable information not existing in the title. This was one of the challenges that occurred while processing a large number of crude oil-based news articles. The authors of [19] proposed a neural network (NN) architecture that quantifies the significance of words and sentences using an inverse similarity scoring metric and transformed them into content and title representations. These representations with the dense layer of NN determined the movements of the stock price and yielded better results when compared against other representations like Bag of Words (BoW), FastText, and structured events [19].

These methodologies remain seldom explored for predicting crude oil prices. According to the Energy Information Administration (EIA) [1], oil prices have an effect on 96% of the transportation industry, 40% of industrial products, and 21% of commercial, and residential applications. In general, higher oil prices affect the economy and can cause inflation [1]. A novel approach introduced in [26] uses sentiment of foreign policy to anticipate the direction of crude oil prices. A drawback of this method is collecting data from a single source (e.g., Twitter), crude oil prices are extremely volatile and are tightly bound with politics and governments across the world. But, it also depends on factors outside of government control, and foreign policies. The ongoing pandemic has caused oil prices to drop below zero for the first time. Therefore, it is necessary to consider data from multiple sources when predicting a highly volatile market. The significance of [26] is the application of the granger-causality test, in this test it determines whether two time-series are correlated.

A convolutional neural network (CNN) approach for text-based crude oil forecasting presented in [18] proposes a method that builds a relationship between the sentiment of news articles and price change. In contrast to the earlier approach, online news articles [18] are analyzed instead of social media due to less noise and a more reliable context. Sentiment analysis and text mining play a significant role in processing unstructured data and making market predictions. The authors of [18] developed a sophisticated approach to identify hidden patterns by extracting features from news headlines, crude oil prices, and financial market data and transform them into a time-series with a CNN and sentiment of news text. A unique benefit of this approach is its capacity to forecast trends of oil price without considering historical data. They also grouped events into political, social, natural catastrophe occurring in the time series using Latent Dirichlet allocation (LDA). Although a deep learning method delivers a tremendous rate of performance using random forest and linear regression, examining only headlines may be misleading and may not present a holistic view of a news article.

The extent to which oil prices and airline delays affect the profitability of an airline industry depends upon an airline's revenue. A comprehensive report detailed in [4] assesses the cost and impact of flight delays in the US. In recovering from the disastrous events in 2001, the capacity increased by 22%. However, during this time, airline delays also increased by 25% as a result of declining aviation infrastructure. A global recession in 2008 helped decrease airline delays but, at the cost of reduced passengers and with a loss of $60 billion in revenue [4]. The above timeline of events describes the aviation industry and its transformation through several global crises, including an unforeseen surge in capacity after recovery. Although the global pandemic takes a toll on the aviation industry in 2020 [23], capacity is anticipated to surge by 30% by 2025, thereby doubling flight delays and a further decline in airline profits. Federal Aviation Administration (FAA) started a project to renovate aging US aviation infrastructure [24] through investment in communication, navigation, surveillance, and automation.

III. Data

The data for this project consists of three steps. The first step involves web scraping news articles related to crude oil prices on [13]. We have decided to web scrape data from [13] as it contains thousands of news articles from 2012. These news articles reflect the latest news about politics, production capacity, foreign sanctions, seasonal consumption, and other events that influence crude oil prices. This website also contains articles that illustrate the effect on oil prices as a result of seasonal demand, social-political issues, natural disasters, and future contracts. It is necessary to extract a broad spectrum of articles as prices are volatile and are dependent on several factors. The news articles are first scraped, cleaned, and ordered by timestamp.


```
"2016-12-30 11:30:00": "Turkmenistan has managed to avert the loss of one of its only two buyers
"2016-12-30 10:41:00": "Coal has been one of the big stories of 2016. With this beat-down market
"2016-12-29 12:05:00": "2017 could be a banner year for oil and gas companies who are looking fo
"2016-12-29 12:01:00": "The OPEC deal might have revived oil prices for now, but a combination o
"2013-06-27 16:33:00": "The British Geological Survey released a report on Thursday that provide
"2012-12-28 18:05:00": "NextEra Energy Resources is the self-proclaimed largest solar power comp
"2012-12-27 18:10:00": "European industry officials and domestic consumers alike are upset with
"2012-12-27 18:09:00": "Chevron, like many international oil and gas companies, plans to develop
"2012-12-27 18:06:00": "Since the last official report released by the IPCC back in 2007, global
"2012-12-27 18:05:00": "Volcanic eruptions can affect the climate. Huge eruptions release millio
```


Fig. 1. Scraped news articles ordered by timestamp

Figure 1 illustrates news articles stored as a key-value pair, where the key represents timestamp, and the value represents the text of the news article. The intuition behind storing it in this format is to process the data as a time series in tandem with crude oil prices. The text articles are first converted to a sentiment value using VADER (Valence Aware Dictionary and Sentiment Analysis) [15] and then plotted on a graph along with the time stamp.

Date	Price	Open	High	Low	Vol.	Change %
16-Sep-19	62.9	61.48	63.38	58.77	1.40M	14.68%
13-Sep-19	54.85	55.15	55.68	54.44	599.47K	-0.44%
12-Sep-19	55.09	55.93	56.34	54	845.98K	-1.18%
11-Sep-19	55.75	57.89	58.3	55.61	858.77K	-2.87%
10-Sep-19	57.4	58.03	58.76	57.2	755.47K	-0.78%
9-Sep-19	57.85	56.8	58.16	56.58	646.99K	2.35%
6-Sep-19	56.52	56.19	56.95	54.83	714.29K	0.39%
5-Sep-19	56.3	55.95	57.76	55.75	712.47K	0.07%
4-Sep-19	56.26	53.92	56.58	53.84	682.30K	4.30%
3-Sep-19	53.94	55	55.24	52.84	970.80K	-1.32%
2-Sep-19	54.66	54.81	55.22	54.36	-	-0.26%
1-Sep-19	54.8	55	55	54.56	-	-0.54%
30-Aug-19	55.1	56.63	56.72	54.55	708.27K	-2.84%
29-Aug-19	56.71	55.88	56.89	55.43	630.76K	1.67%
28-Aug-19	55.78	55.71	56.75	55.34	674.05K	1.55%
27-Aug-19	54.93	53.76	55.72	53.69	596.62K	2.40%
26-Aug-19	53.64	53.25	55.26	52.96	679.02K	-0.98%
23-Aug-19	54.17	55.35	55.6	53.24	807.15K	-2.13%

Fig. 2. A subset of WTI crude oil prices.

The second step involves obtaining data from EIA, the benchmark for these oil prices is known as West Texas Intermediate (WTI) [7]. Followed by applying various forecasting algorithms on this data to determine the trend. Figure 2 describes a subset of WTI crude oil prices containing various attributes such as price, open, high, low, volume generated, and percentage change from the previous reported day. For this project, time series forecasting requires two attributes date and price to determine daily, weekly and monthly trends of crude oil prices.

YEAR	MONT	DAY_OI	DAY_C	OP_CAR	ORIGIN	ORIGIN_CITY	ORIGI	ORIGIN_STA	DEST	DEST_CITY_N	DEST_STAT
2016	1	10	7	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	11	1	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	12	2	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	13	3	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	14	4	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	15	5	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	17	7	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	18	1	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	19	2	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	20	3	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	21	4	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	22	5	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	24	7	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	25	1	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	26	2	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	27	3	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	28	4	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas
2016	1	29	5	AA	SAN	San Diego, C	CA	California	DFW	Dallas/Fort \	Texas

Fig. 3. Airline on-time performance data in 2016

The final step involves extracting data from the U.S Department of Transportation (DoT) [25] as a CSV file. It contains 5 million rows and 43 columns, among which the central attributes are origin, destination, airline code, departure time, arrival time, delay, distance, air time, and carrier code. Figure 3 represents the on-time performance data of all major U.S airlines. Exploratory data analysis revealed missing information in the extracted data, this led to eliminating attributes containing more than 90% missing values. Followed by replacing attributes containing less than 20% with the average and ignoring attributes containing less than 10% missing values.

IV. Methodology

Fuel costs constitute roughly one-third of an airline's operating cost. Consequently, high oil prices hurt airlines' profitability. We have performed a trend analysis of crude oil prices by analyzing the sentiment of large documents obtained through web scraping of news articles. Implemented machine learning algorithms and used forecasting libraries to determine trends of crude oil. In addition to high oil prices, frequent cancellations and delays also impact operating costs. Although some delays are because of air traffic congestion and weather conditions, others are a result of mechanical problems and the airlines' inability to handle capacity. We conducted data analysis on airline performance data to determine the effect of delays on operating costs. The remainder of this section discusses two major factors affecting airline profitability.

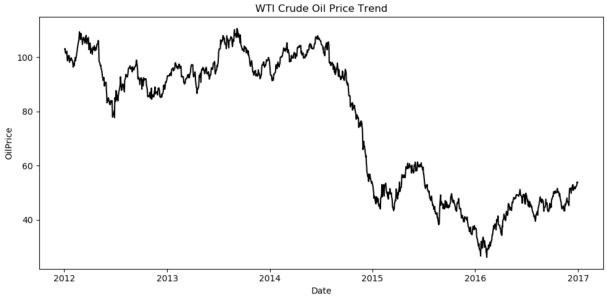

Fig. 4. Historical Crude Oil Price Trend

A. Analyzing historical data

The process began by observing historical crude oil prices to measure the reliability and performance of the model. Figure 4 represents the historical oil price graph. We observed the plunge in crude oil prices during early 2015 and processed the sentiment of news articles during that time frame.

Sentiment analysis plays a crucial role in applications ranging from social media marketing to analyzing patient feedback [28]. It uses text analysis to derive subjective knowledge which can be used to enhance the quality of a product or predict trends by tracking public opinion. In general, it helps to determine the consensus about the item of interest which could either be positive, negative or neutral. The initial step in sentiment analysis requires the identification of objective and subjective components in a text. The former contains factual information and the latter describes sentiments in the form of adjectives [16]. Sentiment analysis can be performed at three levels namely:

- Document: Determining polarity of the entire document.
- Sentence: Each sentence in the document is associated with a polarity.
- Phrase: Analysis and classification of phrases into positive, negative or neutral sentiments.

B. Sentiment of news articles

There are various techniques to perform sentiment classification on text obtained through web scraping. The two main approaches described in a survey [16] are lexicon, knowledge-based, and machine learning approach. Having implemented sentiment analysis using SVM, naïve bayes (NB) and logistic regression to analyze movie reviews in a class project. In this project, we perform sentiment analysis with an unsupervised machine learning algorithm at the document level using VADER. It is a lexicon and rule-based tool to analyze sentiments expressed in social media [15]. The benefit of using VADER is its ability to handle sentences that contain negations, punctuation, slang words, acronyms and emojis. It is built in to the NLTK library of Python with pretrained models that generalize to multiple domains and reduces the complexity of parsing large documents from exponential to linear time [15].

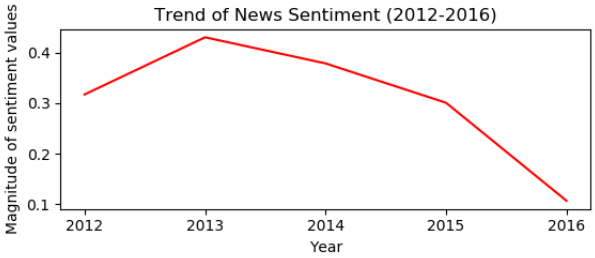

Fig. 5. Sentiment of News Articles.

Figure 5 describes the trend of news sentiments collected over five years between 2012 and 2016. This period was interesting, due to the sudden plunge in oil prices dropping by nearly 54%. The construction of this graph involved ordering thousands of news articles by timestamp, eliminating URLs, and invalid string data followed by computing a compounded sentiment value and averaging them by year. The compounded sentiment value describes the extent to which it is positive or negative. This downward trend of sentiment is closely correlated to the plunge observed in figure 4, this drop boosted airlines earnings and increased profits. After classifying news articles using sentiment analysis, the series of events that resulted in the drop of crude oil prices are visualized by means of a word cloud.

A word cloud is a popular method used to visualize a text corpus [30] to understand frequently used words. In a word cloud, frequently occurring words in the corpus are more highlighted and have a larger font size than words with low frequency. For this project, a word cloud described in figure 6 determines the cause for the change in prices. Some of the words highlighted are 'barrel', 'deal', 'natural gas', 'pipeline', 'China', 'Iran', 'Russia', 'production', 'US', 'OPEC' and 'BP'. The next step resulted in identifying news articles to verify the accuracy of the word cloud. Identifying this information revealed that [21] declining Chinese productions at a rate of 10% each year, slow growth in emerging

Fig. 6. Word Cloud for news articles

markets, several cuts in energy investment, and an imbalance in supply and demand triggered the oil plunge. Over the years, several factors have impacted oil prices including OPEC's supply cuts, U.S sanctions on oil exporters, the health of the global economy, etc., this makes it highly volatile and nearly impossible to forecast. But, integrating data from news articles and observing the trend of historical data lead to better forecasts and help companies to diversify investments and build resilience against fluctuating prices.

C. Impact of high oil prices on airlines

Fuel costs are a highly variable expense for airlines worldwide, constituting between 20-30% of the total operating expenditure. Hence, a shift in crude oil prices impact profitability extensively and cause a ripple effect throughout the airline industry, including airline manufacturers, airport retailing, and airline alliances. To mitigate the uncertainties of varying oil prices and reduce the overall expenses of jet fuel, some airlines hedge fuel prices (e.g., Southwest, Delta, Alaska). Hedging fuel prices is a futures contract that allows a company to establish a fixed cost for some duration, this [33] reduces a company's exposure to fluctuating fuel costs. However, if fuel prices drop, the company remains liable to pay above the market rate thereby, reducing profits. These extended unprofitable periods force businesses to go into bankruptcy. For example, AA filed for bankruptcy [29] after suffering losses for four consecutive years during this time the oil prices also plunged to the lowest level saving AA $4.5 billion at the end of 2015.

The above example describes a situation where low oil prices helped an airline company turn profitable and exit bankruptcy within three years. Hence, fuel hedging does not always work in favor and impacts profitability of an airline company to a large extent. These conditions make the airline industry highly sensitive to small changes in oil prices impacting some companies more than others. For this reason, forecasting is necessary and plays a critical role [6] in planning, expansion and budgeting. By forecasting, companies can make critical business decisions by estimating their expected profits or losses.

D. Forecasting using Prophet

The previous section showed the association connecting the sentiment of news articles and crude prices. This section discusses the implementation of forecasting procedures. Forecasting is always performed by analyses of the past and is an extrapolation of the past into the future. In this project, we make use of an open-source business forecasting tool called Prophet [8], developed by Facebook. Prophet works by modeling a time series, producing forecasts and evaluating them. The accuracy of the model is improved through automated feedback and incorporating domain knowledge [20].

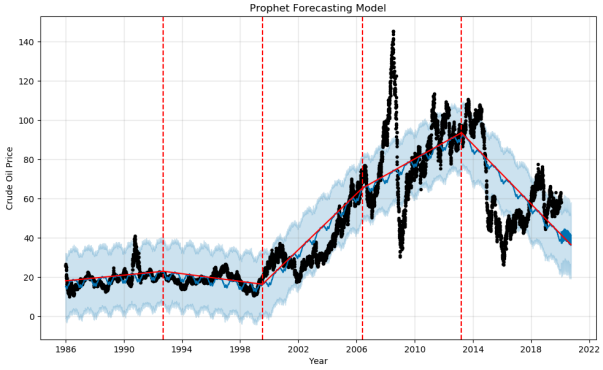

Fig. 7. Forecasting using Prophet

The Prophet framework uses a custom data frame [20] with two columns 'ds' and 'y' to determine the seasonality of a time series data. The column 'ds' consists of date-time objects, and column 'y' consists of crude oil prices. Prophet consists of 'fit' and 'predict' methods for forecasting. The following steps describe the process of building a forecasting model.

In the first step, we choose an arbitrary value that indicates how far to predict in the future and was chosen based on the performance of the forecasting model. Section IV-E discusses the correlation between the error rate and time in more detail. The next step involves implementing 'make_future_dataframe(periods)' method using an arbitrary value for 'periods'. Figure 7 illustrates the forecasting result, on implementing the above steps. The graph consists of the following parts:

- Original Data: The original historical crude oil price data is represented as a black line.
- Forecasting Model: The blue line represents the oil price forecast.
- Change Points: The vertical red dashed lines are referred to as change points and represents the period of changing trends.
- Confidence Interval: The light blue area represents the confidence interval, it determines a range of plausible values for some parameter value. In this case, it represents the confidence interval of oil prices during a period.
- Trend: In case of non-linear time series, seasonality plays an important role for determining weekly, monthly or yearly trends. The solid red line represents the trend of crude oil prices with seasonality removed. This helps to extract seasonality and understand predictive modeling on time series data.

Drawing an inference from figure 7, the downward trend of crude oil prices for the current year is a great indication for airline profitability. However, an unprecedented situation, travel bans, and a sharp drop in air travel prompted the airline industry to seek government bailouts or face bankruptcy [17]. The international air transport association (IATA) projected savings up to $28 billion due to the plunge in crude oil prices. Nonetheless, the organization also predicts a loss of $113 billion in 2020 as a result of travel restrictions and state-wide lockdowns.

E. Metrics for measuring accuracy

Horizon	MSE	RMSE	MAE	MAPE	MDAPE
35 Days	7.401834	2.720631	2.296837	0.120287	0.078196
36 Days	8.359563	2.891291	2.448674	0.128834	0.078832
37 Days	9.445561	3.073363	2.616097	0.138305	0.087169
38 Days	10.649196	3.263311	2.791844	0.148342	0.111755
41 Days	12.016186	3.466437	2.989933	0.159681	0.113495

Fig. 8. Performance Metrics

In time series analysis, we often predict something based on values observed in the past. The accuracy of time series analysis, measures, how well a model performs on a new dataset. The following performance metrics described in figure 8 measure error rate in forecasting: mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and median absolute percentage error (MDAPE). The performance metrics compute the prediction performance as a function of time [8]. Horizon indicates the prediction into the future, a closer look into MSE indicates increasing error rate with time. An experiment conducted to compare the original and predicted values of 2018 resulted in an MAE of 2.59 this value remains consistent with the performance metrics obtained in figure 8. In the next section, we discuss forecast errors and their correlation with time.

F. Measuring forecast error

From the previous section, we can infer that forecast errors increase with time. But, to develop an early warning system, an airline company must know how far they can predict crude oil prices and its accuracy. This forecasting can assist with data-driven decisions including whether to hedge fuel prices or not, saving millions of dollars and ensuring businesses don't enter bankruptcy.

The forecast is evaluated using Prophet's cross validation method along with the following parameters:

- Horizon: Determines how far into the future to predict, we choose the value for the horizon as 365 days.
- Initial: Denotes the size of the training period, in this case, the size of the training period is 1290 days.
- Period: The period is the time between two cutoff dates, period = 180 days.

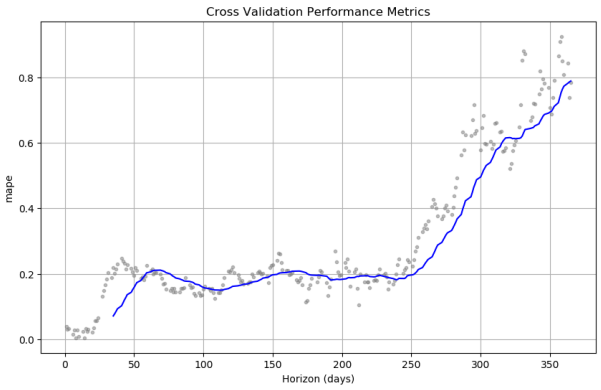

Fig. 9. Performance of forecasting

Figure 9 describes the performance of forecasting, using cross-validation, we plot MAPE with the horizon to determine the integrity of the forecast. It's observed that the forecast accuracy decreases as the forecast horizon expands. The error rate increases from 5% during the first 50 days to 20% for the next 250 days. This value sharply increases and reaches a peak of 80% for 350 days. Hence, this forecasting model performs better for the first 50 days and above average for the next 250 days. The next section discusses another factor affecting the profitability of airlines.

G. Exploring flight delays

Airline on-time performance data includes all domestic flight arrival and departure in the US, between January 1987 to February 2020. It consists of several gigabytes of data with nearly 165 million records. In this project, we extracted and analyzed data between 2016 and 2019 using Python and Tableau.

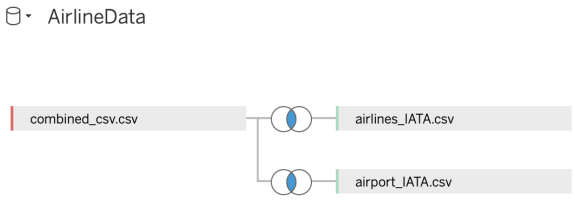

Fig. 10. Data source of flight delays

Figure 10 describes the data sources used for analyzing airline data, it consists of on-time performance data, IATA code associated with airlines and airports. The above data sources are combined using inner join on IATA code and origin city. The US model described in [10] defines the cost of delays as the sum of the cost of fuel and maintenance. Delays also affect the US economy [22] domestic flight delays cost the economy $31.2 billion in 2007. High inefficiencies in the transportation [22] are detrimental for associated businesses and make them less productive.

Aircraft delays are categorized into five categories namely arrival or departure, national airspace system (NAS), security, and weather delay. Figure 11 presents the analysis for the cause of flight delays in 2019. The delays are classified as follows:

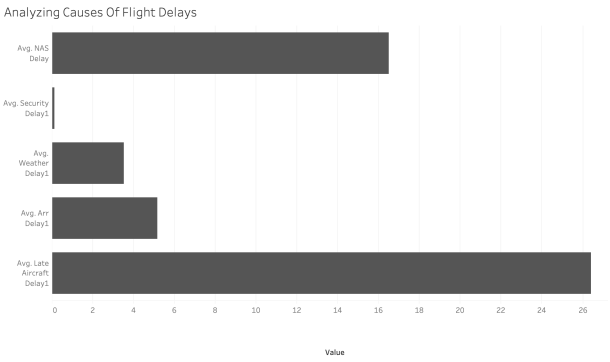

Fig. 11. Cause of flight delays

- Delays due to the air carrier and late arrival contribute to 61%,
- NAS delays account for 32%,
- Weather delays account for only 6%,
- Security delays only account for 1% of the total delays (in minutes).

The above values remain consistent with the analysis described in [31] during the twelve years between 2003 and 2015. The largest contributors to delays are consistently attributed to the late arrival of an aircraft and the aviation infrastructure.

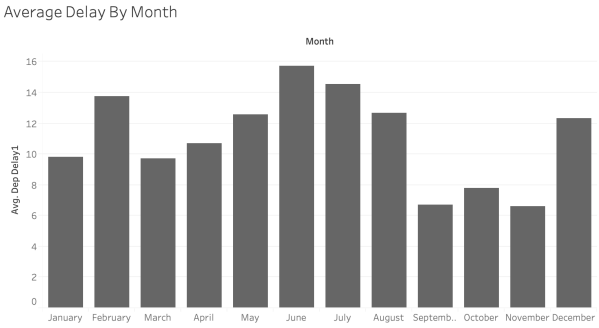

Fig. 12. Monthly average of flight delays

Figure 12 shows the average monthly flight delays during 2019. Drawing an inference from the figure 12, the biggest delays occur during the summer months. Delays are highest during June and July, and also during the peak holiday season of December and February. Since the airline business is seasonal, summer months have high traffic compared to the rest of the year [2]. With higher traffic and declining aviation services, the percentage of delayed flights increases thereby, decreasing profits between two to three months each year.

V. Results

The analysis described in this project can contribute in the form of a prototype that serves to recognize operational

variables that affect the profitability of airlines. While the forecasting model for crude oil price achieved an MAE rate of 2.599 for short term forecasts (e.g., less than 30 days), long-term forecasts (e.g., greater than 150 days) have an error rate of 12.57, and this value increases with time.

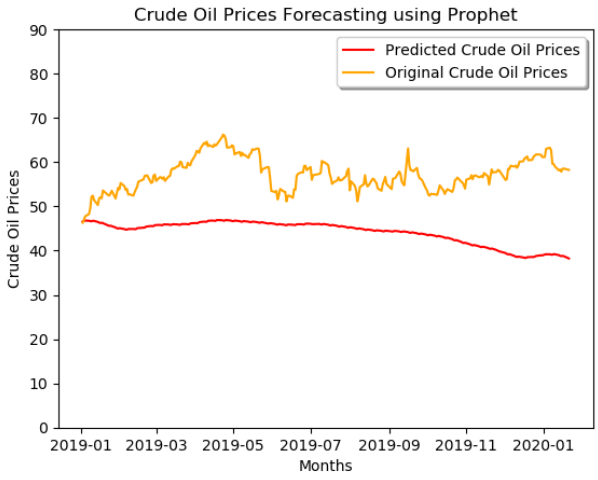

Fig. 13. Predicted and original crude oil prices for 2019-2020

Fig. 13 shows the comparison between predicted and original crude oil prices during 2019-2020. Predicted prices using prophet are lesser than but close to the original crude oil prices. However, due to several fluctuations in crude oil prices in recent months, predicted oil prices have a very high error rate. A limitation of this approach is a higher error rate for forecasts during unusual events. A possible solution is to extract features and analyze sentiment of news articles from several media sources using a CNN classifier.

Delays cost US airlines $62 every minute, and 20% of all domestic flights were delayed with an average of 66 minutes [5]. Moreover, the analysis using Tableau showed "JetBlue" had the highest delays, and "Hawaiian Airlines" had the lowest delays during 2019. These rankings have been consistent since at least 2017 when "JetBlue" lost an estimated $300 million due to delays [5]. A more sophisticated model can be built by analyzing air traffic patterns, determining optimal routes, and upgrading to more advanced aviation infrastructure.

VI. Conclusion and Future Work

Crude oil forecasts for the short term help airline companies decide whether to hedge fuel prices or not. However, a system that incorporates unusual events (e.g., the current massive drop in crude oil prices) and data from multiple sources (e.g., Twitter, news aggregation websites) can lead to more accurate long-term forecasts. A challenging task was scraping thousands of news articles as the script took several days to gather the required data. Hence, in the future, a data pipeline is built to periodically scrape updated data including, oil prices, airline performance data, and news articles to improve forecasts. Another area of growth is exploring the possibility of chunking data from a large file for multiprocessing and further improving time complexity. Also, topic modeling algorithms such as LDA can be used to group features and characterize the effects of various news topics. In conclusion, earlier improvements would lead to more comprehensive results and provide a system where airlines can improve seasonal profitability and capital. These upgrades can benefit airline businesses during an unprecedented time.

Acknowledgment

I would first like to thank my advisor Dr. Carol Romanowski who guided me in the right direction and provided feedback throughout the project. I would also like to extend my gratitude to Prof. Joe Giegel for guiding me through the project and poster presentation.

References

[1] K. Amadeo. Crude oil, its types, uses, and impact, June 2019. [Online; accessed 20-Mar-2020].
[2] AvJobs. Airline economics, August 2020. [Online; accessed 23-Apr-2020].
[3] J. Baffes, M. A. Kose, F. Ohnsorge, and M. Stocker. The great plunge in oil prices: Causes, consequences, and policy responses. *Consequences, and Policy Responses (June 2015)*, 2015.
[4] M. Ball, C. Barnhart, M. Dresner, M. Hansen, K. Neels, A. Odoni, E. Peterson, L. Sherry, A. Trani, and B. Zou. Total delay impact study: a comprehensive assessment of the costs and impacts of flight delay in the united states. 2010.
[5] Bloomberg. Jetblue heads for worst year of flight delays in a decade, December 2017. [Online; accessed 20-Mar-2020].
[6] S. Chand. Business forecasting and its importance to business, March 2016. [Online; accessed 14-Apr-2020].
[7] F. E. Data. Crude oil prices: West texas intermediate (wti) - cushing, oklahoma, August 2020. [Online; accessed 20-Apr-2020].
[8] Facebook. Prophet - python api, March 2018. [Online; accessed 16-Apr-2020].
[9] E. F. Fama, L. Fisher, M. C. Jensen, and R. Roll. The adjustment of stock prices to new information. *International economic review*, 10(1):1–21, 1969.
[10] J. Ferguson, A. Q. Kara, K. Hoffman, and L. Sherry. Estimating domestic us airline cost of delay based on european model. *Transportation Research Part C: Emerging Technologies*, 33:311–323, 2013.
[11] V. Flovik. How (not) to use machine learning for time series forecasting: Avoiding the pitfalls, June 2018. [Online; accessed 24-Apr-2020].
[12] N. Frost. Oil prices are historically low, but it won't help most airlines, March 2020. [Online; accessed 25-Mar-2020].
[13] J. Geiger. Latest energy news, March 2020. [Online; accessed 20-Mar-2020].
[14] W. Horton. Air france-klm faces $1 billion fuel hedging loss as oil price falls due to coronavirus, March 2020. [Online; accessed 24-Apr-2020].
[15] C. Hutto. Vader sentiment, March 2020. [Online; accessed 15-Mar-2020].
[16] H. Kaur, V. Mangat, and Nidhi. A survey of sentiment analysis techniques. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, pages 921–925, Feb 2017.
[17] A. Kimani. Why the oil price crash won't save airlines, March 2020. [Online; accessed 18-Apr-2020].
[18] X. Li, W. Shang, and S. Wang. Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 35(4):1548–1560, 2019.
[19] Q. Liu, X. Cheng, S. Su, and S. Zhu. Hierarchical complementary attention network for predicting stock price movements with news. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 1603–1606, New York, NY, USA, 2018. Association for Computing Machinery.
[20] S. Liu. Forecasting with prophet, December 2018. [Online; accessed 16-Apr-2020].
[21] V. MARC STOCKER, JOHN BAFFESDANA. What triggered the oil price plunge of 2014-2016 and why it failed to deliver an economic impetus in eight charts, March 2020. [Online; accessed 15-Mar-2020].

[22] C. B. Michael Ball. Annual u.s. impact of flight delays, November 2010. [Online; accessed 16-Apr-2020].
[23] K. Miller. This is how flights will change post-coronavirus, April 2020. [Online; accessed 30-Mar-2020].
[24] D. of Transportation. What is nextgen?, August 2019. [Online; accessed 20-Apr-2020].
[25] B. of Transportation Statistics. Reporting carrier on-time performance (1987-present), December 2020. [Online; accessed 16-Jan-2020].
[26] M. Oussalah and A. Zaidi. Forecasting weekly crude oil using twitter sentiment of u.s. foreign policy and oil companies data. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pages 201–208, July 2018.
[27] E. B. Peterson, K. Neels, N. Barczi, and T. Graham. The economic cost of airline flight delay. *Journal of Transport Economics and Policy*, 47(1):107–121, 2013.
[28] V. Ramanathan and T. Meyyappan. Twitter text mining for sentiment analysis on people's feedback about oman tourism. In *2019 4th MEC International Conference on Big Data and Smart City (ICBDSC)*, pages 1–5, Jan 2019.
[29] A. Schmidt. How did american airlines recover from bankruptcy?, March 2016. [Online; accessed 14-Apr-2020].
[30] SumedhKadam. Generating word cloud, March 2020. [Online; accessed 8-Mar-2020].
[31] T. Team. What is the impact of flight delays?, August 2016. [Online; accessed 15-Apr-2020].
[32] T. Team. How sensitive is american airlines' price to crude oil prices?, June 2018. [Online; accessed 15-Apr-2020].
[33] Wikipedia. Fuel hedging, March 2020. [Online; accessed 10-Mar-2020].