The impact of COVID-19 on machine learning models in a commercial aviation use case.

Albin Salmi

Degree Thesis

Information Technology

2022

EXAMENSARBETE	
Arcada	
Utbildningsprogram:	Informationsteknik
Identifikationsnummer:	
Författare:	Albin Salmi
Arbetets namn:	The impact of COVID-19 on machine learning models in a commercial aviation use case.
Handledare (Arcada):	Dennis Biström

Sammandrag:

COVID-19 pandemin har haft kraftiga påverkningar på världen. Maskininlärningsmodeller har blivit negativt påverkade eftersom de använder sig av historisk data för att göra beslut. Förändringarna pandemin har orsakat har lett till att modellernas beslut skiljer sig kraftigt från verkligheten med eventuellt katastrofala följder. Detta beror på att den historiska data som modellerna baserar sig på inte speglar de nya omständigheterna. Flygindustrin är den industri som pandemin påverkat mest och är beroende av maskininlärning. I flygindustrin har modellerna använts för att t.ex. planera tidtabeller och prissätta flygbiljetter. Målet med arbetet var att se om det är möjligt att utveckla modeller som kan förutspå mängden av flygpassagerare, undersöka om det var möjligt att mäta avkastningen av modellernas prestanda mellan år och diskutera möjliga åtgärder. Prognosverktygen Prophet och SARIMAX användes för att utveckla modellerna. Som data för att träna modellerna användes mängden flygpassagerare till och från USA på internationella och inhemska flyg mellan åren 2003-2018. Modellerna användes för att ge en prognos för åren 2019 & 2020. Modellerna klarade av att förutspå året 2019 med en avkastning på endast några procent. Prognosen för 2020 kastade med över 60%. Det var tydligt att modellerna inte kunde förutspå förändringen som pandemin orsakat. Det visade sig att det är möjligt att försöka minska de negativa effekterna på modellerna genom att periodvis utföra mätningar. Om man märker att modellens prestanda försämras kan man ta den ur bruk före den utför dåliga beslut. För att försöka fixa modellen kan man introducera ny data till den då den kan möjligtvis hitta korrelationer mellan ändringarna och därmed anpassa sig.

Nyckelord:	Maskininlärning, Data Drift, COVID-19,Tidsserie prognos, Flygindustri, Säsongsvariation, Prophet, SARIMAX, Regression models
Sidantal:	38
Språk:	Engelska
Datum för godkännande:	

DEGREE THESIS	
Arcada University of Applied Sciences	
Degree Programme:	Information Technology
Identification number:	
Author:	Albin Salmi
Title:	The impact of COVID-19 on machine learning models in a commercial aviation use case.
Supervisor (Arcada):	Dennis Biström

Abstract:
The COVID-19 pandemic has led to unforeseen changes in the world. These changes have had a serious impact on many machine learning models. These models rely on historic data to generate forecasts and make decisions. The changes in society have led to these forecasts and decisions not matching reality, with potentially catastrophic results. The aviation industry is the industry that has been affected the most by the pandemic. Airliners rely on machine learning models to e.g. schedule flights and price tickets. The aim of the thesis was to see if it was possible to develop models capable of predicting airline passenger amounts, research if it was possible to measure the degradation in performance between years and discuss potential ways to fix the models. The forecasting procedures used in developing the models were Prophet and SARIMAX. The data used was the monthly amount of air passengers carried by flights in the USA between the years 2003-2018. The models were then used to forecast the years 2019 & 2020 and then compared to actual numbers. The models were capable of forecasting the year 2019 with a margin of error of only a few percent. When forecasting 2020 the forecast was off by over 60%, making it clear that the models were incapable of adapting to the changed circumstances. There proved to be various ways one could try to fix the model. Active tracking of the models followed by the introduction of new data could lead to the models making connections between old and new data, potentially making the models able to adapt to the new circumstances.

Keywords:	Machine Learning, Data Drift, COVID-19, Time-series Forecasting, Airline industry, Seasonality, Prophet, SARIMAX, Regression models
Number of pages:	38
Language:	Engelska
Date of acceptance:	

CONTENTS

Figures

Tables

1. INTRODUCTION

1.1 Data Drift

The COVID-19 pandemic has had unforeseen effects on the world. The behaviours of governments, individuals and businesses all over the world have been subject to unprecedented changes. All these unpredictable shifts in behaviour have had serious ramifications on machine learning models. These models rely on historic data to detect patterns in order to predict the future. This is not the first time we are faced with a global pandemic, it is, however, the first time we are faced with a pandemic of a society changing magnitude in the era of a data-driven society. Our society relies on these machine learning models to make decisions that allow us our way of life.

Many organizations use machine learning and artificial intelligence systems to aid them in strategic decision-making. The sudden, drastic changes in consumer and corporate behaviour caused by the COVID-19 pandemic has had a major impact on the performance and accuracy of forecasting models. A set of predictions covering a period of time is also known as a forecast in the field of machine learning.

Deshpande (2020), describes a machine learning model as a process that automates decision-making by recognizing patterns that match inputs to outputs, e.g. “if x then y”. A machine learning model requires a data set, referred to as a training set, which is sometimes combined with expert human input. Using the training set, the machine learning model is trained to make accurate decisions. Typically, the size of the dataset will correlate to the accuracy of the model. A portion of the training set is often put aside before training so that the model can be tested on data it has never encountered before, this is called a test set.

These models must be agile enough to handle massive amounts of real-time data. Changes in the environment due to e.g. shifts in consumer preferences, technological innovations and catastrophic events degrade the predictive ability of the models since they are at that point built, trained and tested on data that is no longer relevant to the new, changed circumstances.

Data drift can be described as a variation in the incoming data that is trying to be predicted from the data used to test and validate the model (Machiraju 2021). Concept drift refers to the change in the meaning of incoming data streams and predictions. Neither data nor concept drift are new phenomena in the field of data science. The COVID-19 pandemic has accelerated the rate of concept and data drift aggressively, and they have as a result reached unprecedented levels. A new wave of major drift is anticipated to happen as human behaviour is altered once again as the world begins to prepare for COVID recovery. (Haviv 2020)

Air carriers rely heavily on these machine learning models to do things such as schedule flights, price tickets, predictive maintenance and optimize fuel efficiency (Altexsoft 2021). When these models are exposed to the pandemic caused drift, it could have catastrophic consequences if active efforts are not made to monitor and understand the underlying causes of the changes. If underlying causes are understood, efforts can be made to fix the models.

1.2 Motivation

The COVID-19 has garnered a lot of interest towards drift in machine learning. Drift is a key issue in machine learning because it relies on an assumption that the past equals the future. Unfortunately, in the real world, this is not often the case. Therefore, it is critical to understand the relationship between changing data and model behaviour both before a model is deployed and on an ongoing basis during deployment. During the pandemic, models across various industries were operating in uncharted territory due to the changes in environment and data. (Mardziel 2021)

According to an analysis conducted by Frost & Sullivan (2020), the airline IT market generated a revenue of $21.20 billion by 2019, a big chunk of it in next-generation digital solutions such as machine learning models and AI. Many of these have had to be decommissioned due to the pandemic caused drift. Retraining the models by adding or changing the data has been viewed as a silver bullet, however, retraining without sufficient understanding of the underlying causes and consequences can lead to poor performance (Mardziel 2021). The goal of this thesis is to illustrate and measure the discrepancies caused by the pandemic on self-made models, as well as discuss the causes and problems and plausible solutions.

1.3 Research Questions

In this thesis, I will be trying to answer the following questions:

1. Is it possible to develop simple Machine Learning models capable of giving accurate predictions on flight passenger amounts during normal, non-pandemic circumstances?
2. Is it possible to demonstrate the effects of drift by comparing the forecasts of the models on pre-pandemic and pandemic era data?
3. What can be done about drift?

2. METHODS

2.1 Mapping of the problem

In regular circumstances, machine learning models predict future air passenger volumes with great accuracy. With COVID-19 causing enormous amounts of data drift, no models were able to predict a year-on-year passenger decline of 90% for April 2020. Even though flights started to recover towards the end of the year, the year was still devastating for the aviation industry. Fleets of passenger planes were put into storage, and millions of workers were either furloughed or laid off. (Deshpande 2020).

Knowing the future trend of passengers travelling through air transportation is of immense importance in today's world (Adrangi et al., 2001). According to Carson et al. (2011), it is important to forecast the number of airline passengers as accurately as possible, as such predictions can be used in many contexts, ranging from simple initial planning to complicated business decisions. An analysis conducted by S&P Global (Haydon & Kumar 2020) showed that the industry the pandemic has affected the most is the airline industry, which is why I have chosen it as my target industry.

I am going to be developing two, separate models with the goal of predicting the amounts of passengers on commercial flights for the years 2019 & 2020. I will be using data from 2003-2018 as training data, and then comparing the forecasts of 2019 and 2020 to the real values. Furthermore, I will be measuring the difference in performance between both models for both years by utilizing various performance metrics.

2.2 Data

The dataset used is compiled by the United States Department of Transportation (USA Bureau of Transportation, 2020). The dataset is a CSV file containing the number of passengers carried by both domestic and international flights in the USA between October 2002 and October 2021.

2.2.1 Data preparation

Year	Month	Domestic	International	Total
2003	1	43,032,450	9,726,436	52,758,886
2003	2	41,166,780	8,283,372	49,450,152
2003	3	49,992,700	9,538,653	59,531,353
2003	4	47,033,260	8,309,305	55,342,565
2003	5	49,152,352	8,801,873	57,954,225
2003	6	52,209,516	10,347,900	62,557,416
2003	7	55,810,773	11,705,206	67,515,979
2003	8	53,920,973	11,799,672	65,720,645
2003	9	44,213,408	9,454,647	53,668,055
2003	10	49,944,935	9,608,358	59,553,293
2003	11	47,059,495	9,481,886	56,541,381
2003	12	49,757,124	10,512,547	60,269,671
…	…	…	…	…
2018	12	63,646,582	19,520,179	83,166,761

Table 1. Structure of the dataset. *(USA Bureau of Transportation, 2020)*

Table 1 illustrates the structure of the dataset. In order to prepare the data for forecasting, I combined the month and year columns, removed the commas and chose to only use the values from the “Total” column. The values in the “Total” column were comma-separated, so I removed the commas and transformed the values to the data type integer from the data type string. The columns were then named ds(datestamp) and y, ds containing the dates and y containing numeric values representing the measurement to be forecasted, as shown in Table 2

	ds	y
0	2003-01-01	52 758 886
1	2003-02-01	49 450 152
2	2003-03-01	59 531 353
3	2003-04-01	55 342 565
4	2003-05-01	57 954 225
…	**…**	**…**

Table 2. Dataset ready to be modelled. (USA Bureau of Transportation, 2020)

Due to the fact that I want to compare the forecast of 2019 & 2020 to the actual values, the years after 2018 are left out from the training data and are used as test data for comparing the performance of the models over the years.

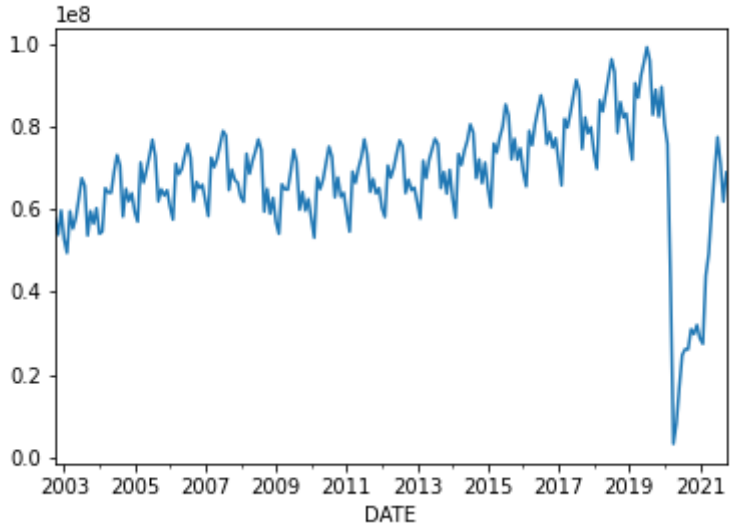


Figure 1. Plotting all the data, illustrating seasonality.

In Figure 1, we can see that every year follows more or less the same structure with a clear spike during summer months and around the Christmas holidays, this means there is seasonality in the data. Seasonality is something to be taken into

account when choosing an appropriate model, not all models are capable of handling seasonality. Additionally, I tested that the data is in fact seasonal by conducting an Augmented Dickey-Fuller test on the data.

2.3 Modeling

2.3.1 Prophet

Prophet is an open-source procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data, making it a perfect match for forecasting flight passenger amounts. (Facebook 2021)

According to Letham & Taylor (2017.), the important idea in Prophet is that by doing a better job of fitting the trend component flexibly, it is able to more accurately model seasonality and the result is a more accurate forecast. Prophet uses a flexible regression model similar to curve-fitting instead of a traditional time series model because it gives it more modeling flexibility, makes it easier to fit models and handles missing data and outliers more gracefully.

Taylor & Letham (2017.), state that the model uses a decomposable time series model that has three main model components: trend, seasonality and holidays. The components are combined in the following equation:

$$y(t) = g(t) + st(t) + h(t) + \epsilon_t$$

Where,

- $g(t)$ = The trend function which models non-periodic changes in the value of the time series.
- $s(t)$ = The periodic changes (e.g. weekly and yearly seasonality)
- $h(t)$ = The effects of holidays that occur on potentially irregular schedules over one or more days.

- ϵ_t = An error term that represents any idiosyncratic changes that are not accommodated by the model.

(Taylor & Letham, 2017.)

Prophet is designed to have easily adjustable parameters that can be changed without knowledge of the underlying model (Taylor & Letham, 2017.). I opted to use ParameterGrid from Scikit-learn which generates a grid of parameters that can then be iterated through to find the best model.

2.3.2 SARIMAX

Hayes (2021.), describes the Seasonal Autoregressive Integrated Moving Average Exogenous model also known as SARIMAX as a version of Autoregressive Integrated Moving Average (ARIMA) which is capable of handling data with seasonal and exogenous factors. Hayes, states that ARIMA is a statistical analysis model that utilizes time-series data to either predict future trends or better understand the data set at hand. According to Hayes, a statistical model is autoregressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future prices based on its past performance or forecast a company's earnings based on past periods.

ARIMA gauges the strength of one dependent variable relative to other changing variables. The goal of the model is to predict the future by examining differences between values in the series instead of through actual values. (Hayes, 2021)

An ARIMA model can be understood by outlining each of its components as follows:

- Autoregression (AR): refers to a model that shows a changing variable that regresses on its own lagged, or prior, values.
- Integrated (I): represents the differencing of raw observations to allow for the time series to become stationary (i.e., data values are replaced by the difference between the data values and the previous values).

- Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

A SARIMAX model has the addition of being able to handle seasonal data (S) and exogenous variables (X).

(Hayes, 2021)

When setting up an ARIMA model, parameters have to be passed to the model in order for it to perform optimally on the data at hand. The parameters can be defined as:

- p : The number of lag observations in the model; also known as the lag order.
- d : The number of times that the raw observations are differenced; also known as the degree of differencing.
- q : The size of the moving average window; also known as the order of the moving average.

(Hayes, 2021)

In addition to the parameters above, a parameter m has to be implemented as well due to the seasonal nature of the data. m indicates the periodicity, i.e. the number of periods in a season, in this case, 12, as the data is monthly.

Due to the seasonal nature of the data, the data is non-stationary. Therefore, the data has to be differenced through the parameter d to make it stationary. According to Hayes (2021), seasonality, or when data shows regular and predictable patterns that repeat over a calendar year, can affect the regression model negatively. If a trend appears and the data is not stationary, many of the computations throughout the process cannot be made efficiently. Isaksson (2020), states that differencing is a technique that deals with trend and seasonality. Differencing stabilizes the mean by removing changes in the level of observations by taking the difference of the observation at hand with the previous one. Figure 2 illustrates

stationarity by comparing the same data pre-differencing and post-differencing.

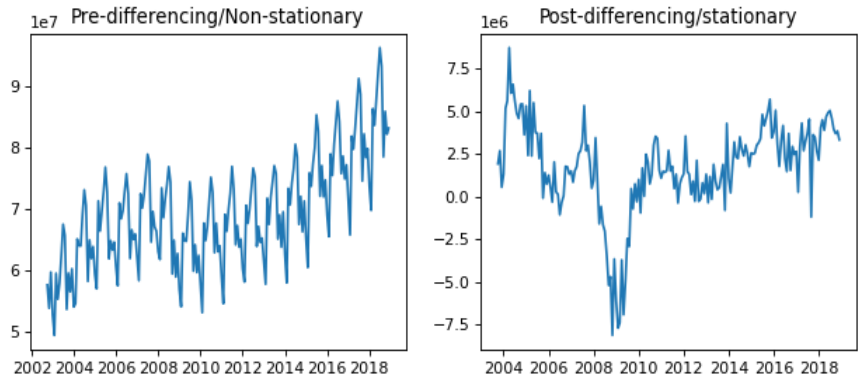


Figure 2. Same data, non-stationary vs. stationary.

To find the optimal parameters for the dataset, I am using the Auto_Arima function from Pmdarima. Auto_Arima does an automatic search to find the best combinations of parameters possible for the given dataset by using a stepwise algorithm to minimize the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). (Rdocumentation, 2022)

$$AIC = 2K - 2\ln(L)$$

Where,

- K: The number of independent variables used
- L: The log-likelihood estimate (a.k.a. the likelihood that the model could have produced your observed y-values).

(Bevans, 2021)

$$BIC = \ln(n)k - \ln(L)$$

Where,

- L is the maximized value of the likelihood function of the model
- n is the number of data points
- k is the number of parameters to be estimated

(Analyttica Datalab, 2021)

2.4 Performance metrics

2.4.1 R-Squared

R-squared (R^2) is a statistical measure that represents the quantity variance for a dependent variable that's explained by independent variables in a regression model. Where correlation explains the strength of the relationship between a dependent and independent variable, R-squared, explains to what extent the variance of one variable explains the variance of the second variable. For example, if the R^2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs. (Fernando, 2021)

$$R^2 = \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

Where,

- y**:** The actual value.
- $\hat{y}$**:** The predicted value of y
- $\bar{y}$**:** The mean value of y values

2.4.2 RMSE

According to Statistics How To (2020.), Root Mean Square Error (RMSE) can be described as the standard deviation of residuals, also known as prediction errors. Residuals measure how far off the data points of the prediction are from the regression line, RMSE measures how spread out the residuals are. RMSE tells us how close the predictions are to the line of best fit. RMSE is commonly used in many fields to verify experimental results.

$$RMSE = \sqrt{\overline{(f-o)^2}}$$

Where,

- f = Values of forecast
- o = Known values

(Statistics How To, 2020)

2.4.3 MAPE

Mean Absolute Percentage Error (MAPE) is a measure of forecast accuracy. MAPE measures accuracy in percentage and is calculated as the average absolute percentage error for each time period minus actual values divided by actual values. In other words, MAPE tells us how far off the forecast was from the actual values in a format that is easily converted to a percentage.

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

Where

- n is the number of fitted points
- A_t is the actual value
- F_t is the forecast value

(Indeed Editorial Team, 2021)

3. RESULTS

3.1 Performance metrics results

3.1.1 Prophet

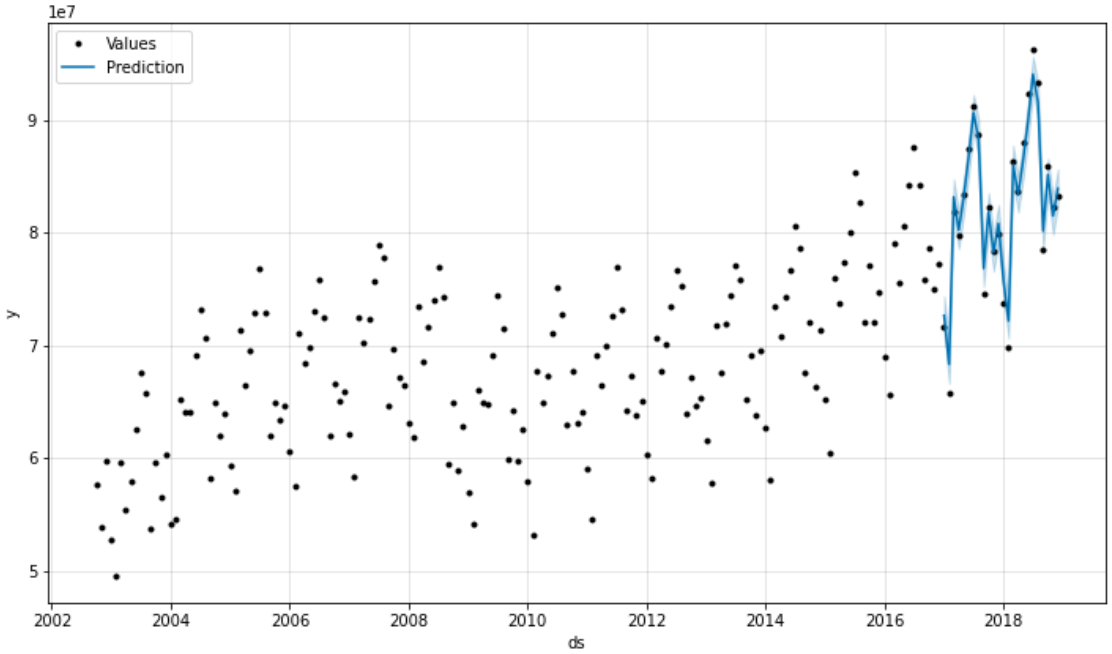


Figure 3. Comparing the forecast from the Prophet model to actual values, pre-pandemic

In Figure 3, we can see that the predictions made on pre-pandemic data are very accurate, especially when taking into account the lower and upper bounds depicted in light blue. The bounds are the maximum and minimum values the model predicts the values can be also known as the uncertainty interval, giving the predictions a bit of a buffer.

Metric	2019 Values	2020 Values	2019 & 2020 Difference
R^2	0.92	-6.93	-6.01
RMSE	3 768 810	63 644 719	59 875 909
MAPE	0.0378	4.7107	4.3327

Table 3. Performance metric results for the Prophet model

The model has an R^2 value of 0.92 measured on test data, meaning that 92% of the variability around the mean of the data can be explained by the model while predicting one year into the future. When calculating the R^2 score for the predictions of the year 2020, we get a score of -6.93. According to Chugh (2020), a negative R^2 score is possible for a Linear Regression Model, where the predictions are worse than just drawing a horizontal line based on previous data, also known as a regression line.

The RMSE value is 3 768 810 measured on test data, meaning that the average deviation of the prediction compared to the regression line is 3 768 810. When calculating the RMSE value for the predictions of 2020, we get an RMSE value of 63 64 4719.

The model has a MAPE value of 0.0378 measured on test data, meaning that the predictions are only about 3.78% off the real values. When calculating the MAPE value for the predictions of 2020, we get a MAPE value of 4.71076, meaning that the predictions were 471.076% off. Comparing the MAPE value of the pandemic era to the pre-pandemic MAPE, we get a difference of 433.276%.

3.1.2 SARIMAX

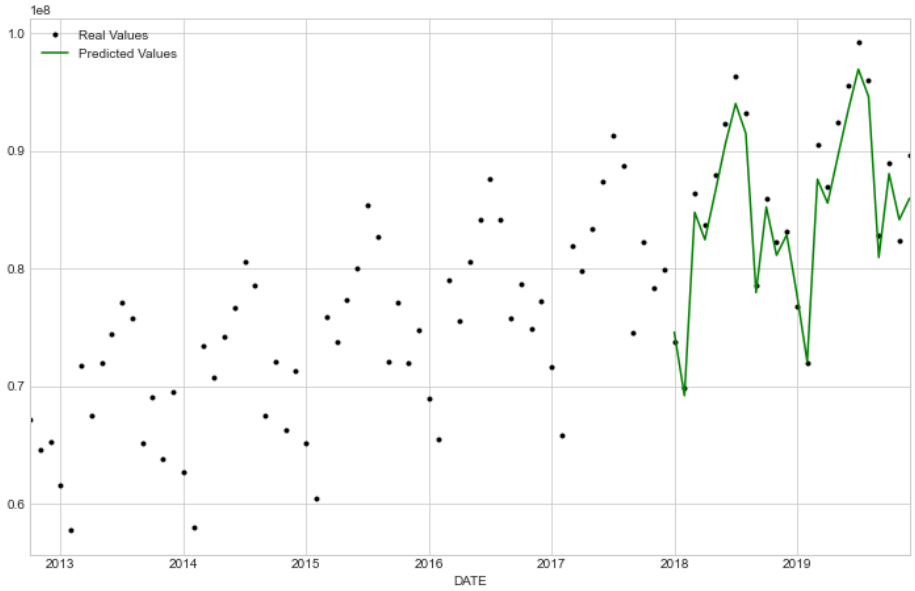


Figure 4. Comparing the forecast to actual values, pre-pandemic.

Metric	**2019**	**2020**	**2019 & 2020 Difference**
R^2	0.94	-6.44	-5.5
RMSE	1 386 715	61 670 524	60 283 809
MAPE	0.0121	4.5700	4.636

Table 4. Performance metric results for the SARIMAX model

The model has an R^2 value of 0.94 measured on test data, meaning that 92% of the variability around the mean of the data can be explained by the model while predicting one year into the future. When calculating the R^2 score for the predictions of the year 2020, we get a score of -6.44.

The RMSE value is 1 386 715 measured on test data, meaning that the average deviation of the prediction compared to the regression line is 1 386 715. When calculating the RMSE value for the predictions of 2020, we get an RMSE value of 61 670 524

The model has a MAPE value of 0.0121 measured on test data, meaning that the predictions are only about 1.21% off the real values. When calculating the MAPE value for the predictions of 2020, we get a MAPE value of 4.57, meaning that the predictions were 457% off. Comparing the pandemic era MAPE value to pre-pandemic MAPE, we get a difference of 444.9%.

3.2 Comparison

Metric	**2019 Prophet**	**2020 Prophet**	**2019 SARIMAX**	**2020 SARIMAX**
R^2	0.92	-6.93	0.94	-6.44
RMSE	3 768 810	63 644 719	1 386 715	61 670 524
MAPE	0.0378	4.7107	0.0121	4.5700

Table 5. Comparison of performance metrics between the two models.

Going through the results and comparing the two models, we can see that the SARIMAX model performs slightly better across the board. SARIMAX has an R^2 score that is 0.02 higher than Prophet, meaning that it should give 2% more accurate predictions. Comparing the RMSE & MAPE scores, the SARIMAX model has a MAPE score of 2.57% lower than the Prophet model, meaning that the predictions are 2.57% closer to the actual values.

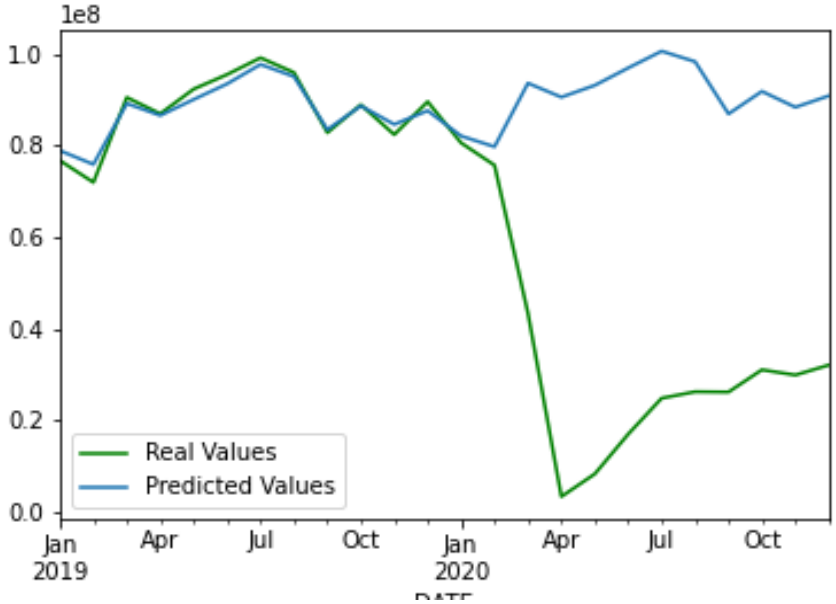


Figure 5. A plot of predictions versus actual values from the Prophet model.

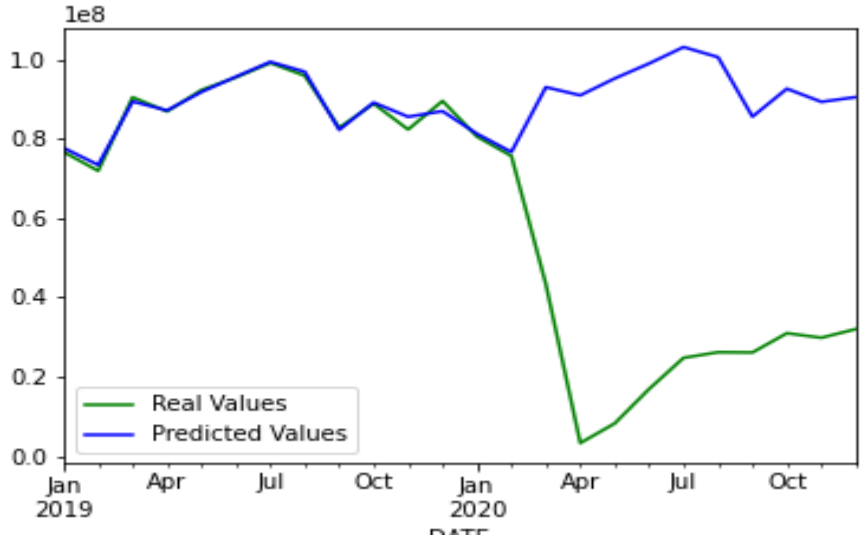


Figure 6. A plot of predictions versus actual values from the Sarimax model.

As expected, neither of the models were able to predict the sudden decrease of passengers caused by the pandemic, as seen in Figures 2 & 3. The SARIMAX model was able to predict the slight dip in January 2020, making its performance metric scores slightly better for the year.

According to the US Bureau of Transportation (2021), the decrease in international air passengers was down approximately 60% overall in 2020. If we then calculate the real decrease in passengers for the years 2019 and 2020 using our dataset, we get a decrease of 62.2%. If we compare the forecast of 2020 to the actual values for 2020, we get a difference of 63.4% for the Prophet model and 62.2% for the SARIMAX model. Neither of the forecasts deviated less from 2019 numbers than the actual data did, indicating that they did not adjust. The models went from giving accurate predictions in 2019 to giving predictions that were significantly worse than the regression line.

Source of values	**Decrease in passengers 2019 to 2020**
Dataset (Reality)	~60%
Prophet	63.4% (Forecast)
SARIMAX	62.2% (Forecast)

Table 6. Comparison between the real decrease in passengers vs. the decrease between the forecasts and reality.

4. ANALYSIS

4.1 Causality

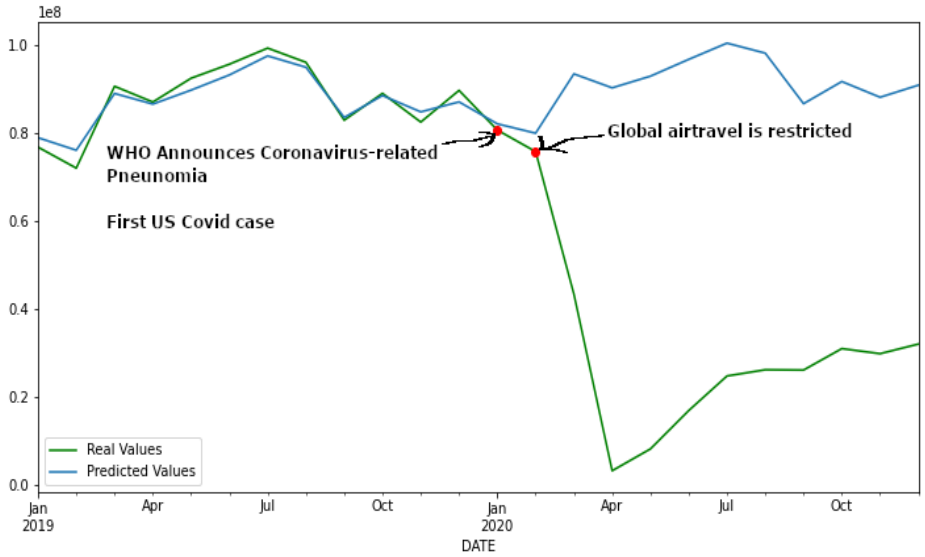


Figure 7. The exact points where real values started to differ from the forecast (Prophet) in a meaningful way.

Examining the timeline (AJMC, 2021) of the COVID-19 developments of 2020 we can pinpoint the exact events that lead to first a slight, but noticeable discrepancy between the model and real data followed by a more substantial one. In January, the WHO announced that Coronavirus-related pneumonia was spotted in Wuhan, China. At this stage, experts already showed concern and advised for travel precautions. Later on, in January, the virus had been spotted outside of China, leading to some airports beginning to screen passengers. Just a day after screening started, the first case was spotted in the USA. In early February, Global air travel was restricted, and the USA declared the outbreak a public health emergency. The events in January were heavily reported across media and are likely the reasons for first the small dip. The massive drop in February was a result of the imposed travel restrictions. The events unfolded quickly, and even a model utilizing real-time data streaming would most likely not have been able to foresee the rapid change in trend.

Airline ticket prices were sharply reduced during 2020 as a response to the pandemic, in an effort to keep the industry alive. In the third quarter of 2020 U.S. domestic flights averaged $244.79, the lowest they have been in more than 25 years, according to the U.S. Department of Transportation (Josephs, L, 2021). The drop in price was not enough to keep up to keep the numbers up to pre-pandemic levels as seen in Figure 8, even though some people started to travel more due to the cheap prices. Airlines also implemented a reduced passenger capacity in order to keep safety distances in order to stop the spread of the pandemic, resulting in a reduced capability of sales.

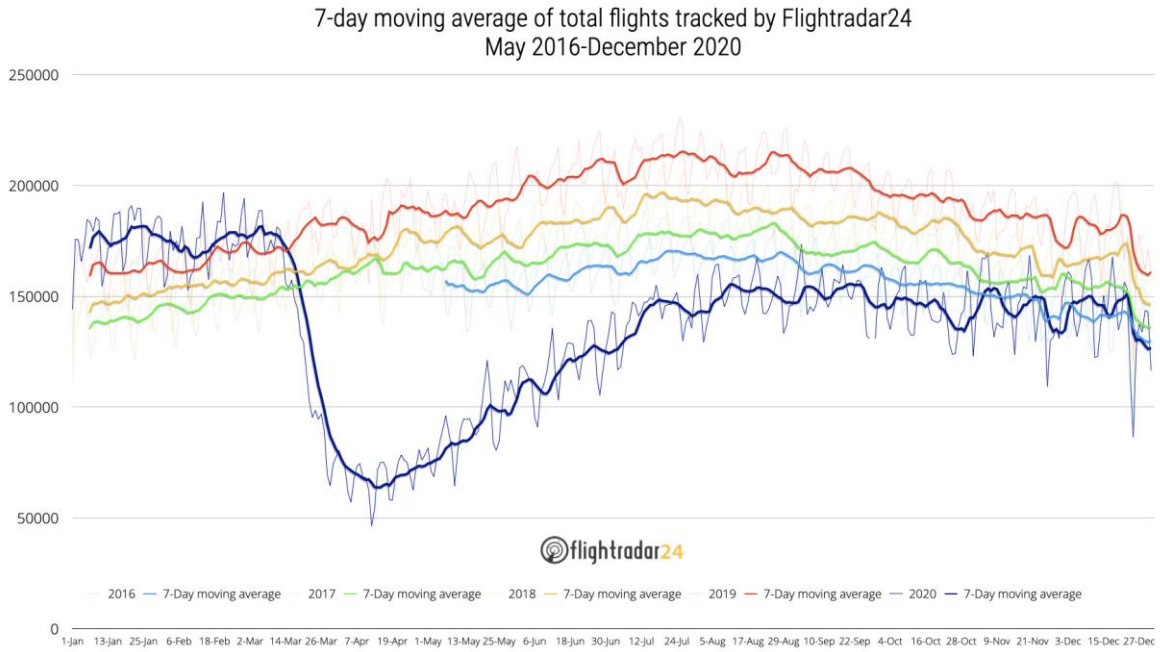


Figure 8. Yearly comparison of the total amount of flights. Petchenik (2021)

Comparing the number of passengers in 2020 to other years in Figure 8, we can see that it looks very similar to the comparison of 2020 vs. the forecasts of the predictive models. This further proves that the models would've been perfectly fine during normal circumstances, but had no way of adapting to the sudden change in trend.

4.2 Conclusion

In conclusion, machine learning models were trained to predict the amounts of the monthly number of passengers carried by both international flights in the USA. Both models were capable of very accurate forecasts for the year 2019. When forecasting 2020, the models were not able to predict the decrease in passengers caused by the pandemic, resulting in an inaccurate, unusable forecast. Analysing the performance metrics, it was evident that the models were severely crippled by the COVID-19 caused data drift. Comparing MAPE values, the Prophet model performed 433.276% worse for the year 2020 while the SARIMAX model performed 463.600% worse compared to 2019. The decrease in passengers between 2019 and 2020 was 62.2%. When comparing it to the differences between the forecasts of 2020 and the actual values for the year, the differences were 63.4% for the Prophet model and 62.2% for the SARIMAX model, making it evident that the models did not foresee the incoming rapid decline in passengers.

Even though the models developed were highly accurate during typical circumstances, they were unable to provide useful forecasts for the atypical year of 2020. These discrepancies between the forecasts and actual data indicate that there has been a significant amount of drift caused by the pandemic in a very short time.

Attempts to mitigate the effects of drift can be made by incorporating different measures. According to Burkhardt et al. (2020) the first method one should implement is actively tracking drift through things such as sequential analysis, model-based methods, or time distribution methods. When drift is then detected, there are a number of ways one could try to make sense of the drift. These ways include things such as: incorporating new data sources, using existing data in different ways, increasing the frequency of data collecting and processing, and acquiring external data to feed into the models. A combination of these measures could help the models make sense of the changes by finding patterns and correlations between the old and new data. According to Isaksson (2020), periodically retraining models with a portion of historical data that better

represents the new circumstances combined with differencing can help alleviate the effects of drift.

5. DISCUSSION

5.1 Plausible solutions

After measuring drift among the predictions of the models, the question arises, what can be done about it?

The first step would be to actively track data drift in deployed models. Machiraju (2021) states that Data drifts can be identified using sequential analysis methods, model-based methods, and time distribution-based methods. Sequential analysis methods like DDM (drift detection method)/EDDM (early DDM) rely on error rate to identify the drift detection, a model-based method uses a custom model to identify drift, whereas distribution-based methods use methods that calculate statistical distance in order to calculate drift between probability distributions. Some examples of statistical methods to calculate the difference between populations are: Population Stability Index, Kullback-Leiber or KL Divergence, Jenson-Shannon or JS Divergence, Kolmogorov-Smirnov Test, Wasserstein Metric or Earth Mover Distance. While actively monitoring drift, one could track degradation of the model and make decisions, such as taking it offline and attempting to rebuild it before the model does any bad decisions.

After detecting drift, the next step would be to expand the models' data sources. According to Burkhardt et al. (2020) whether a model needs to be rebuilt or retuned, there's a considerable chance that a model affected by drift needs new inputs to provide more accurate insights into the period of volatility. Even though we are living in unprecedented times, Burkhardt et al. found that there exists a lot of data that can help better understand current trends, they recommend expanding data sources in the following ways:

- Incorporation of new or previously unused data sources. A North American bank, which had been using traditional credit scores to assess the credit risk of customers, began analyzing account data to uncover gaps in direct deposits or receipts of unemployment in real-time.
- Using existing data in new ways. A manufacturer of automotive engines began leveraging telematics data from its engines in order to improve understanding of traffic patterns when cities began to reopen and to generate demand forecasts. Before the renewal, the telematics data was mostly used in support of maintenance and warranty work.
- Increasing the frequency of data collection and processing. The challenge does not always lie in what data you're collecting and processing, but also in the frequency in which you do so. Shortening learning cycles enables models to become more flexible and adapt faster as events occur in society.
- Acquiring external data. Acquisition of external data can help build a more holistic model, capable of linking the external values to the values it's trying to predict. Such data can come in the form of, e.g. COVID-19 data in order to link the increase or decrease of cases to the prediction being made.

(Burkhardt et al., 2020)

According to Isaksson (2020), in addition to the methods described above, one could implement a so-called "sliding window". A sliding window means that the model is periodically retrained with a small portion of historical data that better represents the new circumstances. Isaksson also states that differencing can be used to handle drift since it helps negate the effects of trend. In the case of our models, differencing did not help. The change was so sudden that the addition of a sliding window to perform differencing upon would only have been possible to implement after the fact.

5.2 Further research

Regarding the addition of external data, it would be interesting to explore the effects of adding features such as monthly COVID cases into the models. Would the models be able to find a correlation between the amount of active COVID cases and the decrease in passengers? The data could be misleading because during 2020, after the rapid decline in passengers happened, both active COVID cases and flight passenger amounts increased for the remainder of the year. An option to work around the mismatch in data would be to feed the model the derivative of cases between months. The models could then potentially see that a rapid surge in cases indicates a new wave of COVID and, therefore, fewer passengers.

As the world starts to shift towards a post-pandemic state, a new wave of drift is already underway. The removal of pandemic-related restrictions is once changing society in unpredictable ways. The world is not expected to revert to exactly how it was before the pandemic started, meaning that models trained on pre-pandemic data will not necessarily perform as well as they did before. First, models had to be adapted to the changed circumstances of the pandemic era, but when things start to change again, we are faced with the challenge of adapting them to a post-pandemic era.

REFERENCES

Adrangi, B., Chatrath, A. & Raffiee, K., 2001. *The demand of US air transport service: a chaos and nonlinearity investigation.* Transportation Research, Part E, pp.337-53

AJMC Staff, 2021. *A Timeline of COVID-19 Developments in 2020,*

Retrieved 15.02.2022

From: https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

Analyttica Datalab, 2019. *What is Bayesian Information Criterion (BIC)?,*

Retrieved 15.02.2022

From: https://medium.com/@analyttica/what-is-bayesian-information-criterion-bic-b3396a894be6

auto.arima: Fit best ARIMA model to univariate time series 2020,

Retrieved 15.02.2022

From:

https://www.rdocumentation.org/packages/forecast/versions/8.16/topics/auto.arima

Bevans, R., 2021. *Akaike Information Criterion* | *When & How to Use It,*

Retrieved 15.02.2022

From: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

Carson, R.T., Cenesizolu, T. & Parker, R., 2011. *Aggregate demand for USA commercial air travel*. Int. J. Forst., pp.923-41.

COVID-19 and the aviation industry: Impact and policy responses,

Retrieved 04.02.2022

From: https://www.oecd.org/coronavirus/policy-responses/covid-19-and-the-aviation-industry-impact-and-policy-responses-26d521c1/

Burkhardt et al., 2020. *Leadership's role in fixing the analytics models that COVID-19 broke,*

Retrieved 17.02.2022

From: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/leaderships-role-in-fixing-the-analytics-models-that-covid-19-broke

Deshpande, P., 2020, *How COVID-19 Has Infected AI Models,*

DominoDataLab,

Retrieved 20.01.2022

From: https://www.dominodatalab.com/blog/how-covid-19-has-infected-ai-models

Facebook, 2022. *Prophet: Automatic Forecasting Procedure*

Retrieved 17.02.2022

From: https://github.com/facebook/prophet

Fernando, J., 2021. *R-Squared,*

Retrieved 14.02.2022

From: https://www.investopedia.com/terms/r/r-squared.asp

Haviv, Y., 2020, *Concept Drift and the Impact of COVID-19 on Data Science*,

Retrieved 04.02.2022

From: https://www.iguazio.com/blog/concept-drift-and-the-impact-of-covid-19-on-data-science

Haviv, Y., 2020, *Concept Drift and the Impact of COVID-19 on Data Science*,

Retrieved 04.02.2022

From: https://www.iguazio.com/blog/concept-drift-and-the-impact-of-covid-19-on-data-science

Haydon, D., Kumar, N. 2021, *Industries Most and Least Impacted by COVID19 from a Probability of Default Perspective September 2020 Update*,

Retrieved 04.02.2022

From: https://www.spglobal.com/marketintelligence/en/news-insights/blog/industries-most-and-least-impacted-by-covid19-from-a-probability-of-default-perspective-september-2020-update

Hayes, A., 2021. *Autoregressive Integrated Moving Average (ARIMA)*

Retrieved 10.02.2022

From: https://www.investopedia.com/terms/a/autoregressive-integrated-moving-average-arima.asp

Indeed Editorial Team, 2021. *What Is MAPE? (Plus How To Calculate MAPE in 3 Steps)*,

Retrieved 14.02.2022

From: https://www.indeed.com/career-advice/career-development/what-is-mape

Isaksson, C, 2020. *The Impact of Coronavirus on Machine Learning Models,*

Retrieved 20.02.2022

From: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/leaderships-role-in-fixing-the-analytics-models-that-covid-19-broke

Letham, B., Taylor, SJ. 2017, *Prophet: forecasting at scale*,

Retrieved 09.02.2022

From: https://research.facebook.com/blog/2017/02/prophet-forecasting-at-scale/

Mardziel, P., 2021, *Drift in Machine Learning*,

Retrieved 04.02.2022

From: https://towardsdatascience.com/drift-in-machine-learning-e49df46803a

Petchenik, I, 2021 *Commercial flights down 42% in 2020,*

Retrieved 17.02.2022

From: https://www.flightradar24.com/blog/commercial-flights-down-42-in-2020/

RMSE: Root Mean Square Error. 2020,

Retrieved 14.02.2022

From: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/

USA Bureau of Transportation, 2020. *Passengers All Carriers - All Airports Dataset*

Retrieved 07.02.2022

From: https://www.transtats.bts.gov/Data_Elements.aspx?Data=1

USA Bureau of Transportation, 2021. *Full Year 2020 and December 2020 U.S. Airline Traffic Data*

Retrieved 17.02.2022

From: https://www.bts.gov/newsroom/full-year-2020-and-december-2020-us-airline-traffic-data

APPENDICES

APPENDIX 1. ABSTRACT IN SWEDISH

Introduktion

COVID-19 har orsakat stora förändringar på både individ- och samhällsnivå. Förändringarna har även haft en kraftig inverkan på maskininlärningsmodeller. Dessa modeller använder sig av historisk data för att förutse framtiden. Det är inte förstå gången mänskligheten drabbats av en global pandemi, men det är förstå gången mänskligheten drabbats av en pandemi så kraftig som denna i ett samhälle som drivs av data. Vårt samhälle är beroende av maskininlärningsmodeller som fattar beslut och dessa beslut bidrar sedan till den levnadsstandard vi har idag.

En maskininlärningsmodell kan beskrivas som en process som automatiserar beslutsfattning genom att känna igen mönster i data. Maskininlärningsmodellen kräver ett dataset, som även kan kallas för träningsdata. Denna träningsdata används sedan för att träna upp en modell att göra träffsäkra beslut genom att se mönster i data. COVID-19 har lett till att verkligheten skiljer signifikant från den data modellerna tränats med. Denna avvikelse kallas för “data drift”.

Flygbolag använder sig av maskininlärningsmodeller för att t.ex. planera tidtabeller, prissätta flygbiljetter, förutsäga underhåll och optimera mängden bränsle som går åt. När dessa modeller blir utsatta för “data drift” orsakad av en pandemi eller andra stora förändringar kan detta ha katastrofala följder. För att undvika problem måste en aktiv insats göras genom att övervaka modellerna och sedan försöka förstå de bakomliggande orsakerna till förändringarna. Efter man förstått vad förändringarna beror på kan man försöka åtgärda dem.

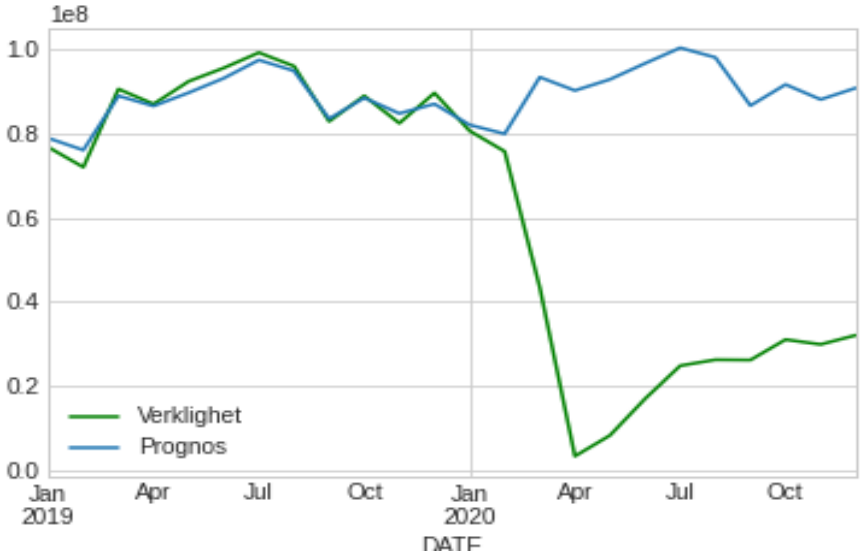
Metoder

Under normala omständigheter används maskinlärningsmodeller för att förutse mängden flygpassagerare. Inga modeller kunde förutse en minskning på 90% från April 2019 till April 2020. Flygindustrin är den industri som pandemin påverkat mest och därför jag valt att fokusera på den i mitt examensarbete.

Jag har utvecklat två modeller som förutser passagerarmängderna för åren 2019 & 2020. Som träningsdata använde jag mig av den månatliga mängden internationella flygpassagerare från och till USA under åren 2003-2018. Denna data innehåller säsongsvariation eftersom mängden flygpassagerare fluktuerar beroende på högtider och årstider. För att utveckla modellerna använde jag mig av prognosverktygen Prophet och SARIMAX. Både Prophet och SARIMAX lämpar sig bra för data med säsongsvariation. Modellernas prognos jämfördes sedan med de riktiga mängderna för 2019 & 2020. För att ge en mer grundlig inblick på skillnaderna mellan prognos och verklighet använde jag mig av olika prestandamätningar som mätte modellernas noggrannhet.

Resultat

Båda modellerna klarade av att förutspå passagerarmängderna för 2019 med hög noggrannhet. MAPE är ett prestandamått som beskriver hur många procent prognosen avviker från verkligheten. För år 2019 avvek Prophets prognos med 3.78% medan SARIMAX prognos avvek med 1.21%. När jag sedan räknade ut MAPE värdet för prognoserna av 2020 avvek Prophet med 47.1076% och SARIMAX med 45.7000%. Sedan räknade jag ut skillnaden mellan MAPE värdena av 2019 och 2020 för att få reda på hur med hur många procent modellens noggrannhet försämrades. Prophet blev 43.3276% sämre medan SARIMAX försämrades med 44.4900%. I följande graf kan man tydligt se skillnaden mellan de två åren.



Analys

Då man jämförde antalet passagerare som modellen förväntade sig och det verkliga antalet märkte man en liten skillnad i januari 2020 och sedan en stor skillnad i februari 2020. Skillnaderna kan förklaras med COVID händelseförloppet 2020. I januari meddelade WHO om att en Coronavirus-relaterad lunginflammation hade upptäckts i Wuhan, China. Experter rekommendera snabbt försiktighetsåtgärder gällande resande. I början av Februari blev flygresor begränsade på globalnivå. En blandning av begränsningar och medians inverkan orsakade den kraftiga minskningen i passagerarantal. Händelserna skedde snabbt och ingen maskininlärningsmodell kunde förutspå vad som skulle komma.

Slutsats

Modellerna som jag utvecklade klarade av att förutspå året 2019 med hög noggrannhet. När det kom till året 2020 var modellernas prognoser värdelösa eftersom de inte alls klarade av att anpassa sig till de förändrade omständigheterna. Antalet flygpassagerare minskade med 62.2% mellan 2019 och 2020. Om vi jämför modellernas prognoser för 2020 med verkligheten får vi en skillnad på 63.4% (Prophet) och 62.2% (SARIMAX). Eftersom värdena är ytterst lika med minskningen mellan 2019 och 2020 kan man säga att modellerna

inte kunde anpassa sig. Detta tyder på att det har uppstått en märkvärdig mängd med “data drift”.

Diskussion

Efter demonstrationen på hur drift kan se ut uppstod frågan, vad kan man göra åt saken?

Första steget är att aktivt spåra drift genom att periodvis utföra mätningar. Då man aktivt söker efter drift kan man göra beslut som att försöka fixa modellen eller ta den ur bruk före den tar ogynnsamma beslut.

För att försöka fixa modellen kan man t.ex. försöka introducera ny data till modellen, använda nuvarande data på nya sätt, höja frekvensen av data insamling och bearbetning eller använda utomstående data. Ny och utomstående data kan användas för att hjälpa modellen se mönster mellan den nya och gamla data och därmed bli bättre på att anpassa sig. Ett exempel på detta kunde vara att mata in mängden skillnaden mellan aktiva koronafall mellan tidsperioder för att se kopplingen mellan storleken av ökning och avtagande av fall och i flygpassagerare.

En ny våg av “data drift” håller redan på att ske då världen återhämtar sig från pandemin. Världen förväntas inte återgå till hur den såg ut före pandemin och detta innebär att modeller som tränats före pandemin inte nödvändigtvis kommer att fungera fastän pandemin når sitt slut.