Predicting air travel demand using LSTM and SARIMAX

Degree Project in Computer Science and Engineering DA231X

Carina Wickström
carinawi@kth.se

1. Introduction

1.1 BACKGROUND

Using machine learning to predict behavioral patterns is widely used in science these days. This thesis uses machine learning to predict air travel demand. We will compare a SARIMAX and LSTM model with time series data for air travel. Other than a strong seasonal pattern in the data, the models will consider a set of different additional features, including: media investment, temperature, covid, news, stock market swings, and exchange rates.

The science of predicting travel demand has been of scientific interest for different applications, and using different techniques [1, 2, 3]. Application areas include targeted marketing [4], as well on a governmental level to support aviation budget planning [5]. The air travel demand also has a direct financial impact on airlines and the tourism sector, hotels, etc. Some research has been made on analyzing which factors impact air travel demand, including national economic factors [6, 7], the impact of AI and robotics [8], marketing [9], and covid [10]. Although the prediction of air travel demand is of clear importance, taking into account the effect of the fast-varying outside factors such as temperature and news, is hardly explored in science today.

We intend to use a SARIMAX model and an LSTM model to predict air travel demand. We chose SARIMAX because it handles seasonality well and has proven to have high performance in a similar context [11]. We chose LSTM because it has also proven to be well suited for prediction based on time series data. It has the additional benefit of supporting multivariate data and catching unknown time lags between important time events.

1.2 Problem

The company (Flygresor) is interested in predicting the amount of customers they will receive in the future based on prior data. The problem is that we don't have a predictive model for air travel demand, and we don't know for which features a SARIMAX model works better or worse than an LSTM model.

1.3 Purpose

The purpose of this thesis is to compare an LSTM model with an SARIMAX model. We will learn for which features one model works better than the other, and draw conclusions regarding which factors affect air travel demand using ML.

1.4 Goal

The goal is to provide a comparison in performance of SARIMAX and LSTM. We will also gain an insight into which features affect air travel. This knowledge can be used by the company for marketing purposes. For example, if the demand on air travel is expected to be low, it might be wasteful to invest in click-ads at that time.

1.5 Benefits, Ethics and Sustainability

Comparative studies between ML models are beneficial for science. It helps us choose relevant models for different use cases, as there exist an abundance of machine learning models to choose among.

The data used in this thesis regards air travel. From an environmental sustainability standpoint, it promotes the efficiency of marketing air travel which has a negative effect on environmental sustainability.

Regarding the long-lasting sustainability of the model itself, one main factor that is taken into account in the model is the effect of covid. The data was separated into pre-covid and post-covid because the data between these times differed greatly. Hence, as post-covid behavior is expected to go back to normal, it is assumed that the pre-covid model will also be relevant after the pandemic is over. One risk with that assumption is that there are lasting behavioral changes that are not taken into account, affecting the accuracy of the model in the future.

Regarding ethics, if due to any inaccuracy the report would make false conclusions regarding people's air travel behavior, this could lead to company loss or confusion for future research.

1.6 Methodology

The work consisted of a prestudy followed by a set of experiments. We built a base model for both SARIMAX and LSTM, using seasonality and marketing investment as model features. Thereafter, we added a set of additional features in sequence that we believe to add further predictive power on the model. The considered features were: temperature, covid, news, stock market swings, and exchange rates. Those features were implemented one by one, and the model was continuously evaluated. From these results, we kept (list) features that were included in a final model.

1.7 Stakeholders

A bit of information about Flygresor

1.8 Limitations

We have only considered LSTM and SARIMAX for the models. The code uses library dependencies, pmdarima[http://alkaline-ml.com/pmdarima/] for the SARIMAX implementations, and keras[https://keras.io/] for the LSTM model. Regarding the data, the base model relies on media investment data which spans from 2016-01-01 to 2021-12-01. The data was split into pre-covid and covid, with the delimiting date being 2020-02-20. Due to this split, the data during covid spans little less than two years, which was a limiting factor in finding long patterns such as yearly seasonality for training the model.

1.9 Outline

The thesis contains the following content:

Section 2: Theoretical background. This section describes the background knowledge that the experiments were built upon. It covers …
Section 3: Method.
Section 4: Results

2. Theoretical background

This section covers background knowledge of series models SARIMAX and LSTM, as well as current literature on the topic of comparing these models, and using machine learning in air travel prediction.

2.1 Time series models

There exist many models that can be used for time series forecasts. We will consider SARIMAX and LSTM.

2.1.2 SARIMAX

SARIMAX (Seasonal Autoregressive Integrated Moving Average with eXogenous factors) is a model used for forecasting time series data. The S indicates seasonality. The AR indicates that the variable is regressed on its own lagged values. The word auto refers to the fact that it's based on previous values of itself. The I indicates that means that the data values are replaced with the difference with the current and previous value. The MA refers to the regression error being a linear combination of previous terms. Finally, the X addition to the model allows for additional feature inputs. As exogenous variables, one can add Fourier terms[x] in order to enable the model to handle multiple seasonalities.[x]
[https://medium.com/intive-developers/forecasting-time-series-with-multiple-seasonalities-using-tbats-in-python-398a00ac0e8a][https://tanzu.vmware.com/content/blog/forecasting-time-series-data-with-multiple-seasonal-periods]

[Math formula of the arima variables here]

2.1.3 LSTM

LSTM (Long Short Term Memory) is a neural network model used in time series. It is an RNN (Recurrent Neural Network) which means that it reuses the output from a previous step as an input for the next step. The node has an internal state as a working memory to store information over many time steps.

For calculation, the node uses the input value, previous output, and the node's internal state.
LSTMs have gates that regulate the saved state's impact on the calculations, how much the state is updated, and the state's influence on the output.

LSTM were developed to deal with the vanishing gradient problem that is a problem with traditional RNNs. An advantage with LSTM compared to other ML models is that they are especially good at handling data even when there exist unknown duration lags between important events in the time series. [https://en.wikipedia.org/wiki/Long_short-term_memory]

[Explanation of the structure of an LSTM cell here]

2.1.3 Comparison

There have been several comparative studies between ARIMA (including its versions) and LSTM.

LSTM, being a more modern approach to time series forecasting. In several cases, it has shown to outperform traditional methods such as ARIMA. [https://www.mdpi.com/2073-4433/11/4/316] In a study from 2018 where ARIMA and LSTM were tested on forecasting financial time series, the study reported that LSTM outperformed ARIMA, as LSTM reduced error rates by 84-87 percent compared to ARIMA.
[https://ieeexplore.ieee.org/abstract/document/8614252?casa_token=0G8C2MKd1LgAAAAA:vbogdugBjF4STicxYejfCNdjpKsPsNSzwQ1UTFZnfztqUXzAeHBZ44NcYejjVNAZm_tl6lbV9KM]

However, there have been contradictory results where ARIMA provides higher accuracy than LSTM for time series forecasting. [https://dl.acm.org/doi/abs/10.1145/3377713.3377722]

Differences in performances could be caused due to differences in the input data. A comparative study from 2020 found that SARIMA performed better than LSTM on long term predictions, whereas LSTM performed poorly on short term predictions.
[https://arxiv.org/abs/2007.08092]

Moreover, LSTM handles data with unknown time lags well, in contrast to ARIMA. Another advantage that LSTM has over SARIMA is that LSTM can fit non-linear data. This way, it can detect seasonalities that were not specified to the model beforehand. LSTM is hence more robust, whereas SARIMA relies on prior knowledge about seasonality. [https://arxiv.org/abs/2007.08092]

Clearly, both LSTM and ARIMA have their benefits. Due to this, there has even been developed a hybrid model SARIMAX-LSTM which attempts to utilize both models. Experiments from one study concluded that the SARIMAX-LSTM model performed with better forecasting effect compared to both SARIMAX and LSTM respectively.
[https://ieeexplore.ieee.org/abstract/document/9233117?casa_token=uFqHO7Ji43AAAAAA:UjjYdouR_W0k4H2ZKQFkdtw2PEalvQoP6TqwvimIyAnGd3LIqceiPHydoIDUxn-VzT5yuQbvmqI]

2.2 Air travel prediction using ML

[The existing work that has been done in predicting air travel]

3. Method

The method consisted of a prestudy phase, which consisted of researching machine learning models and and previous publications. Thereafter, datasets were gathered for the features that should serve as inputs to the models. Then, a programming phase was conducted where the experiments were carried out. Lastly, the results were evaluated.

3.1 prestudy

The prestudy phase consisted of researching LSTM and SARIMAX. For information gathering, Google and Google scholar were used. The prestudy concluded that LSTM and SARIMAX were suitable for time series forecasting.

3.2 data gathering

The work included finding suitable datasets as model inputs. The datasets that were found are the following:

(list the datasets).
(for each dataset, describe how I preprocessed the data)

3.2.1 Aggregation of media investment data

The media investment data was aggregated into three sets, where each set was used as an individual model input feature.

[explain the aggregations]

The aggregation was created based on staff experience from the company from working with the data, with help from a budget planning tool. [explain the tool here]

3.2.2 Calculating temperature deviations

[Explain the complicated procedure that SMHI does for calculating temperature differences]

3.3 programming implementation

- technical stuff

- how we evaluated the model performance by calculating the model accuracy using MAE, MAPE and RMSE

- The Fourier terms are a combination of sinusoids that have periods of 365.25 days (one year). With this additional feature, we are able to express an additional yearly seasonality.

Libraries
Dependencies

3.3 Evaluation

Describe how I did the final model accuracy compared to the base model using the modified paired student's t-test and calculating p value

4. Results

For each feature, for both pre-covid and post-covid, we trained a SARIMAX and LSTM model. These are the results.

4.1 Base model

We calculated baseline models with SARIMAX and LSTM predicting air travel demand using only seasonality and media investment

How we divided the media investment data

We experimentally tried different lookbacks
The lookback can be determined by looking at the acf plot

What can we read from the correlation graph (ARIMA)

Plots:

LSTM	SARIMAX
Graph	Graph

Accuracy (do a grid)

LSTM:	SARIMAX
MAE=	
RMSE=	
MAPE=	

4.2 Base model + temperature

Weather forecast (from SMHI opendata API), explain the API
Explain how the data is calculated

4.3 Base model + covid numbers

Number of covid death- or infected cases, and stay-at-home-restrictions (datasets gathered from ourworldindata.org/explorers/coronavirus-data-explorer and ourworldindata.org/explorers/covid-stay-home-restrictions)

4.4 Base model + news

Negative or positive news on omni news aggregator regarding covid or air travel (sentiment analysis), considering their effect within a time after its publication

Intuitively rank them

Build a dictionary of words that are positive and negative towards travel.

Scan through the publication and count the words (n-level) that are negative towards travel and positive towards travel respectively. Catch phrases such as “increase covid restrictions”, “flight ban”, “stay home”, etc.

Give each publication a score of travelability, and associate it with the date of publication

4.5 Base model + stock market swings

Stock market swings (dataset from nasdaq.com/market-activity)

4.6 Base model + exchange rate

Currency exchange rate SEK to USD and Euro. Theory: if SEK is strengthened then people are more eager to fly. “it's too expensive to travel there now” - reasoning

4.7 Political events

It has been shown that political events affect the air travel demand negatively.
Show how all website traffic died at the time of the Åhlens crash and the start of the Ukraine-Russia invasion.

4.8 Final model

We will use the modified paired student's t-test proposed method by Nadeau and Bengio to compare the two models and calculate a p value for statistical significance.

Base model SARIMAX vs base final model SARIMAX
Base model LSTM vs final model LSTM

5. Analysis

5.1 Expected result

Hypothesis: adding the new features improves the predictive power of the base model. Our null hypothesis is that the models with the base features, and the models with additional features, perform with the same accuracy. We used the modified student's p test.

6. Conclusion

7. Future work