Predicting 2023/24 Hotel Industry Revenue for the City of Lisbon

PDS: Programming for Data Science

Efe Türkselçi 58418

Luke Gorman 59059

Mark O'Shea 58930

Table of Contents

Abstract:

The tourism Industry in Portugal made up for a total of 11.8 percent of Portugals GDP in 2019. Hotel revenue makes up a significant portion of this value and unsurprisingly the capital city of Lisbon dominates the nations hotel industry revenue. In this present study we predicted the hotel revenue for the city of Lisbon in 2023 and 2024. Seasonal autoregressive integrated moving average (SARIMA) models are commonly used in forecasting and to great effect. Unfortunately, the Covid-19 pandemic has made forecasting any hospitality data rather difficult. To overcome this, we have adopted Meta's Prophet model to sidestep the extreme effect of the pandemic. We were able to reduce our mean absolute percentage error by 50% with hyperparameter tuning and cross validation. The results suggest to us that hotels in the city Lisbon can expect to see revenues of 1.49 bn in 2023 and 1.35 bn in 2024. The prophet model should be considered for time series forecasting in hospitality and tourism data due to its ability to handle extreme events.

Keywords: SARIMA, Prophet, Forecasting, Lisbon, Hotel Industry, Incomes, COVID-19

Introduction:

Recently, tourism has emerged as a vital driver of the economy in Lisbon, with millions of visitors flocking to the city annually. Hotels are pivotal to the tourism industry in Lisbon and knowing their incomes may be leveraged to obtain new investment for expansion which in turn can have positive impacts on the tourism sector for Portugal. In 2019 according to government documents the tourism sector accounted for 11.8% of Portugal's GDP making it a highly significant industry.

Therefore, accurately forecasting tourism revenue has become a critical task for businesses and policymakers alike, enabling them to make informed decisions concerning resource allocation, marketing strategies, and infrastructure investments. The hotel industry in Lisbon is ever expanding. With a compounded annual growth rate (CAGR) of 4.9% for overnight stays in Lisbon this industry can be attractive for investors. Hotels themselves saw a CAGR of 5.1% from the years 2015 – 2018.

The dataset used in the present study can be found on Instituto Nacional De Estatistica (INE), an open database for all statistics associated with Portugal. Data preparation and cleaning was handled with Python. The real challenge was to accurately forecast 2 years in the future and to sidestep the effect of covid lockdowns.

Seasonal Autoregressive Integrated Moving Average SARIMA models like ones used by Chang et al. (2012) [1] are used to predict future values for seasonal data. ARIMA and SARIMA models are traditional approaches that have been popularly applied in the field of meteorology, engineering, and economics while Prophet is a relatively new model that has shown promising results in handling extreme events such as lockdowns caused by the Covid-19 pandemic. In this context, this paper aims to provide a comparative analysis of SARIMA and Prophet models for predicting hotel revenue in Lisbon and to discuss the strengths and limitations of each approach.

The aim of this project is to provide accurate forecasts of hotel revenue for the city of Lisbon in 2023 and 2024. The accuracy will be determined based on mean absolute percentage errors of our constructed models. Once the models have been built, we will deploy them to forecast the next couple of years.

This project and its findings are intended to provide investors and portfolio managers with accurate forecasts such that they can make informed decisions with regards to Investment in Lisbon. Moreover, we can consider city officials and even local government as beneficiaries of these findings. With the insight used to justify investment in local amenities and public transport. Also, stakeholders in the Hotel Industry will find these results encouraging.

Literature review:

In recent years, time series forecasting models, such as SARIMA models, have been widely used for revenue forecasting in the hotel industry. However, the outbreak of Covid-19 in 2020 has brought new challenges to this field, as traditional forecasting models may not be suitable for capturing the drastic changes in hotel revenue patterns caused by the pandemic. There has been little work done on forecasting specifically Hotel Revenue in Lisbon. Nevertheless, there has been plenty of research done on SARIMA models and Prophet models for forecasting seasonal monthly data.

Feng et al. [2] Compared the accuracy of a SARIMA model versus that of Prophet for predicting road traffic injuries in Northeast China. In this study their models were trained on test sets from 2015 to 2019 and accuracy was assessed comparing test data year of 2020 to model predictions for that year. Mean absolute percentage error, root mean squared error and mean absolute errors were used to measure model performance. Prophet was chosen because of its ability to handle highly seasonal data. In this study of highly seasonal and monthly data Prophet outperformed SARIMA.

Overall, while traditional forecasting models such as SARIMA can be useful for predicting hotel revenue in Lisbon, they may not be suitable for capturing the impact of extreme events such as the Covid-19 pandemic. More advanced models such as Prophet may offer better performance in these scenarios, but they also have their limitations. Therefore, a combination of different forecasting models and approaches may be necessary for accurate revenue forecasting in the hotel industry during uncertain times.

In 2020, Santos and Oliveira Moreira [3] reviewed the state of Portugal's Tourism industry post covid 19. The important results of the study and relevant research on the subject are summarized in the literature review that follows. Effect of COVID-19 on Portugal's Tourism Travel bans, lockdowns, and other precautions taken in response to the COVID-19 epidemic have had a substantial negative impact on Portugal's tourist sector, as to be expected. In 2020, there were 76% fewer international visitors to Portugal than the previous year, which had an impact on the country's tourism-related earnings.

Methodology:

The methodology of this project is heavily influenced by the CRISP-DM method [4]. It follows a logical path where business understanding is the initial phase. Our business case was simple. To provide users with accurate forecasts of revenue for Lisbon's Hotel Industry in 2023 and 2024. Naturally this would lead us to gather data from INE which has monthly data on Lisbon Revenue from 2006 to Present. Prepping our data was uniquely less labour intensive as most data science projects. Modelling is the next step to follow, and here we modelled both ARIMA and Prophet models before deployment. Accuracy measures were generated as a means of model evaluation, and Prophet had a much better score than SARIMA. Finally, we could deploy our models to forecast for 2023 and 2024. Below discusses these methods in more detail.

Data

Data was obtained from INE which is an open database for used and updated by the Portuguese government. This database allowed us to obtain monthly hotel revenue in the city of Lisbon for the years 2006 – 2023 totalling 205 entries.

Data Preparation

After the necessary cleaning and preprocessing operations are performed on the data, it is converted into a time series. The time series is plotted to show the original data and this gives us a visual representation of the data. Decompostion is performed to analyze the trend elements of time series.

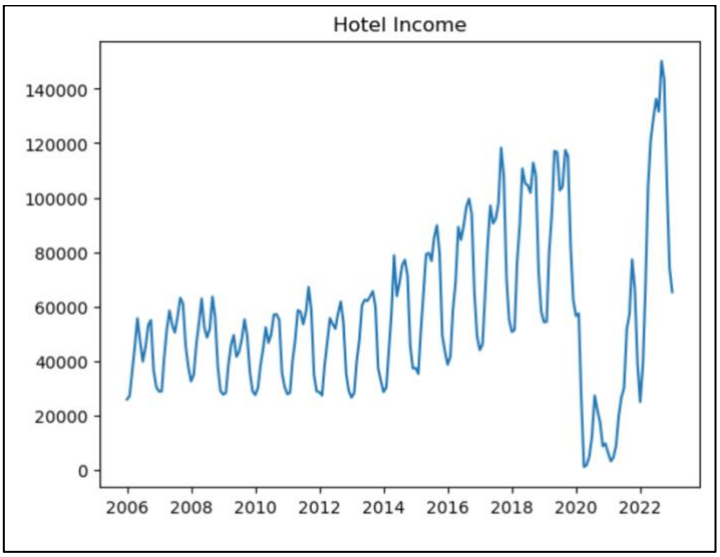

Fig. 1 *shows the general trend of our time series*

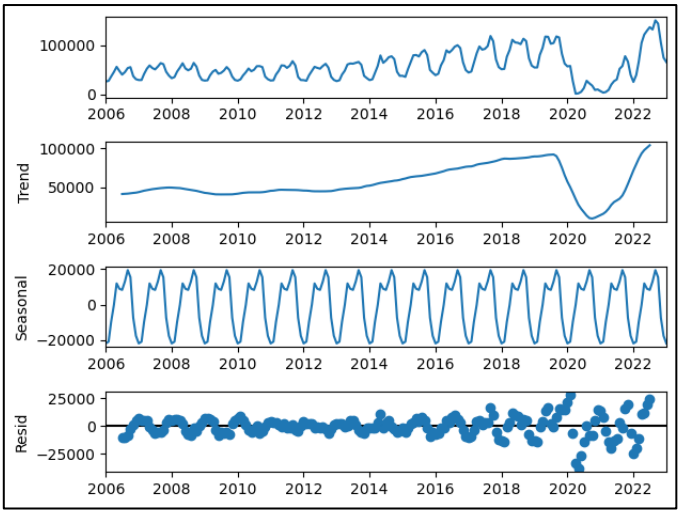

Fig. 2 *Shows the decomposition of the time series. Top shows the actual time series, Trend represents the overall trend in the series, Seasonal refers the seasonal aspect of the series and residuals is what remains of the series once the trend and seasonality are removed.*

MAE/MAPE: The mean absolute error and mean absolute percentage error were the two metrics we decided to base our model's accuracy off. This is obtained when we have a test set and prediction values and the skearn mean_absolute_percentage_error() function is used. These are similar performance metrics used by Feng et al. [2]

ARIMA Model

The general formula for a seasonal ARIMA model is:

SARIMA(p,d,q)(P, D, Q)

The autoregressive and moving average elements are represented by the p and q steps. The d term is the level of differencing required to make the data stationary (see below). The larger PDQ values are the seasonal equivalent of these elements. For example, if we perform seasonal differencing once the D = 1.

Stationarity

Before constructing an ARIMA model one must check that their data is stationary. Stationarity means that a time series exhibits a statistically constant behaviour over time. That is, when the mean, variance, and autocorrelation (self-similarity) of the time series do not change over time, the time series is considered stationary.

```
ADF Statistic: -3.4978903817013864
p-value: 0.00804143911080711
Critical Values:
   1%: -3.4652439354133255
   5%: -2.8768752281673717
   10%: -2.574944653739612
```

Fig. 3 *Results from our Ad-Fuller test, showing that our data is stationary.*

Many functions have been developed in python and R to check for stationarity. One such is the Ad-Fuller test (ADF). Ad-fuller test is a test based on hypothesis testing used to determine the stationarity property of the time series. The Ad-fuller test evaluates whether the time series has a unit root problem. The unit root means that the time series is not stationary. The Ad-fuller test evaluates the stationarity property of the time series by determining whether the time series is a unit root or not and provides more reliable results if the time series satisfies the stationarity condition. In this test the p-value rejects the null hypothesis. When the p value is below 0.05, we can confirm that our data is stationary.

```
 ARIMA(1,0,1)(1,0,1)[12] intercept   : AIC=3630.850, Time=0.16 sec
 ARIMA(1,0,1)(2,0,2)[12] intercept   : AIC=inf, Time=1.02 sec
 ARIMA(1,0,1)(2,0,1)[12] intercept   : AIC=3591.953, Time=0.92 sec
 ARIMA(1,0,1)(2,0,0)[12] intercept   : AIC=3628.330, Time=0.41 sec
 ARIMA(1,0,1)(1,0,0)[12] intercept   : AIC=3628.888, Time=0.40 sec
 ARIMA(0,0,1)(2,0,1)[12] intercept   : AIC=inf, Time=0.87 sec
 ARIMA(1,0,0)(2,0,1)[12] intercept   : AIC=inf, Time=0.94 sec
 ARIMA(2,0,1)(2,0,1)[12] intercept   : AIC=inf, Time=1.16 sec
 ARIMA(1,0,2)(2,0,1)[12] intercept   : AIC=3625.121, Time=0.76 sec
 ARIMA(0,0,0)(2,0,1)[12] intercept   : AIC=inf, Time=0.87 sec
 ARIMA(0,0,2)(2,0,1)[12] intercept   : AIC=inf, Time=0.84 sec
 ARIMA(2,0,0)(2,0,1)[12] intercept   : AIC=inf, Time=1.04 sec
 ARIMA(2,0,2)(2,0,1)[12] intercept   : AIC=3626.112, Time=1.20 sec
 ARIMA(1,0,1)(2,0,1)[12]             : AIC=inf, Time=0.80 sec

Best model:  ARIMA(1,0,1)(2,0,1)[12] intercept
Total fit time: 15.887 seconds
```

Fig. 4 *Screenshot taken from Juptyer notebook showing the auto_arima function iterate though all combinations of models and calculating each AIC score.*

Auto/Partial-correlation functions

Autocorrelation functions (ACF) and partial autocorrelation functions (PACF) are both important metrics in determining our ARIMA model. They will help us determine our q and p terms in our model. They ACF measures the amount of linear dependence between observations in our time series. If the time series is highly autocorrelated, a larger number of lags may be needed to properly capture the behaviour of the data. Conversely, if the time series is less autocorrelated, fewer lags may be needed.

Model construction and fitting

Once this information was acquired, we used a function called auto_arima to iterate though all possible SARIMA models. The model that was selected was a SARIMA (1, 0, 1), (2, 0, 1), [12]. Where [12] represents yearly data. This model was selected based on it returning the lowest AIC score.

This ideal model is then fitted to our timeseries. For model accuracy and performance, we split the data into test and training set. By convention it is suggested that we split the data with an 85/15 percent split. We trained our model on the training set and then checked its performance versus the test set. With our prediction values and actual values, we can calculate our models mean squared error and mean square percentage error.

Forecasting

The Statsmodels package allows for us to predict into the future with the predict function. We first fit our model to our data and then predict 24 months into the future. This returns a new list of values which we can visualize and sum to get the predicted tourism revenue for the years 2023 and 2024.

Prophet Model Building

Prophet is a forecasting model developed by Meta that is designed to handle time series data with strong seasonality and holidays. It is an additive regression model that decomposes time series data into trend, seasonality, and holiday components [2]. Prophet is known for its simplicity and ease of use, making it a popular choice for data analysts. We opted to use the prophet model because of its ability to sidestep the pandemic. With prophet we can assign so called black-swan events or events of extreme nature as holidays. By doing this the model will take less weight in the pandemic period and allow us for a more accurate forecast. For consistency we split the data with an 85/15 split again.

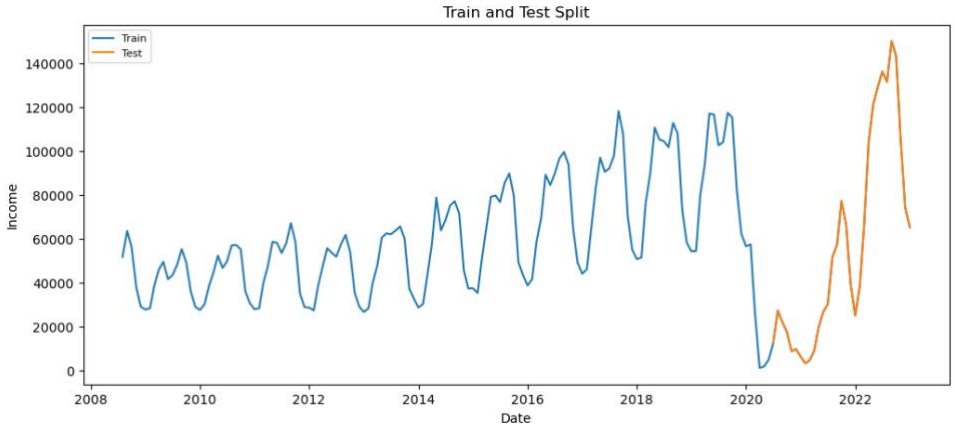

Fig. 5 *Train test split we adopted with the conventional 85/15 ratio*.

The prophet model was trained on the training data and again the predicted values were measured against the actual training data with the holidays (covid lockdowns) passed as a parameter in the model.

Prophet parameters

There are many parameters for the prophet model and because of this we performed a grid search for choosing the best hyperparameters. These parameters were parameters which minimised the MAPE. Some of the other key parameters tuned include:

Changepoints: changepoints are abrupt changes in the time series trajectory. Prophet will automatically detect these changepoints and adapt appropriately. *Changepoint_prior_scale:* which determines how flexible the trend is; the default value is 0.05 and increasing this will allow the trend to fit more closely to the observed data. *Seasonality mode:* which determines which type of seasonality the model uses. For our model, we had an additive seasonality

Cross validation in prophet:

Prophet includes a function for time series cross validation to measure forecast error using historical data. This is done by selecting cut off points in the history, and for each of them fitting the model using data only up to that cut off point. We can then compare the forecasted values to the actual values. This allows us to compute MAPE and Mean Absolute Errors for our tuned model.

Prophet Forecast

Once we built our optimal prophet model with a lower MAPE we could make forecasts for the years 2023 and 2024. Much like the SARIMA forecast we get a list returned which we can visualize.

Results:

SARIMA:

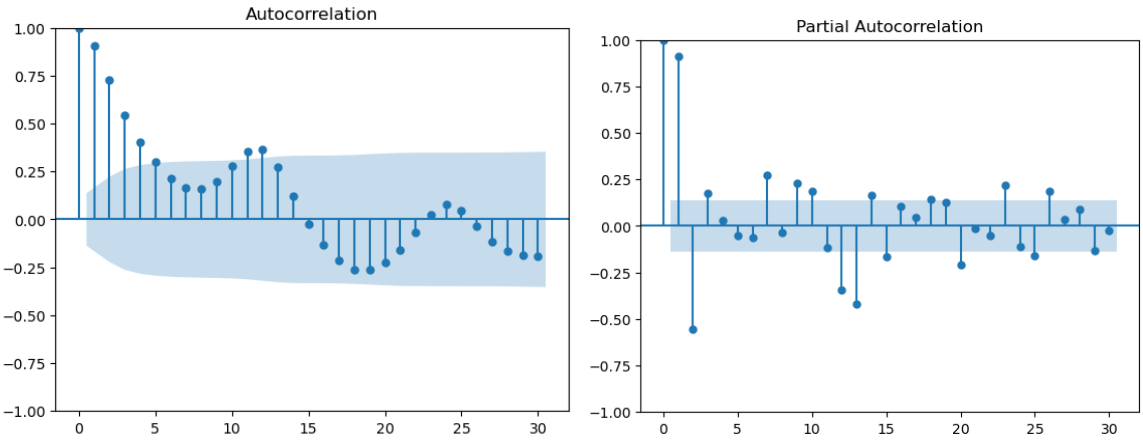

Fig. 6 *AFC and PAFC results from our time series data. For the PACF there is a spike at the beginning and at the 12-month mark which hints at seasonality.*

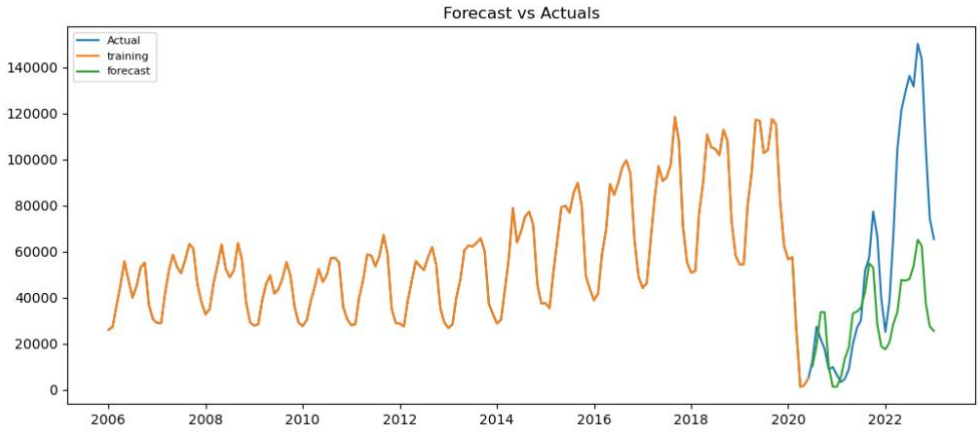

Fig. 7 *Plot of our forecasted values versus the actual values of the time series.*

In figure 7 we can see the limitations of using the SARIMA model with the covid pandemic. Our model had trained on the data before covid but the sharp decrease in revenue at the start of 2020 heavily impacted the model's prediction. As stated in the methodology above we generated predicted values (green) and were able to calculate the mean absolute percentage error when comparing them to the actual test values (blue). For this model we obtained a MAPE of 46% which was very high. Nevertheless, we still forecast the next 24 months (Figure 8) in our time series to use as a comparison versus the Prophet model. It was clear that covid had an extremely negative effect on the model's ability

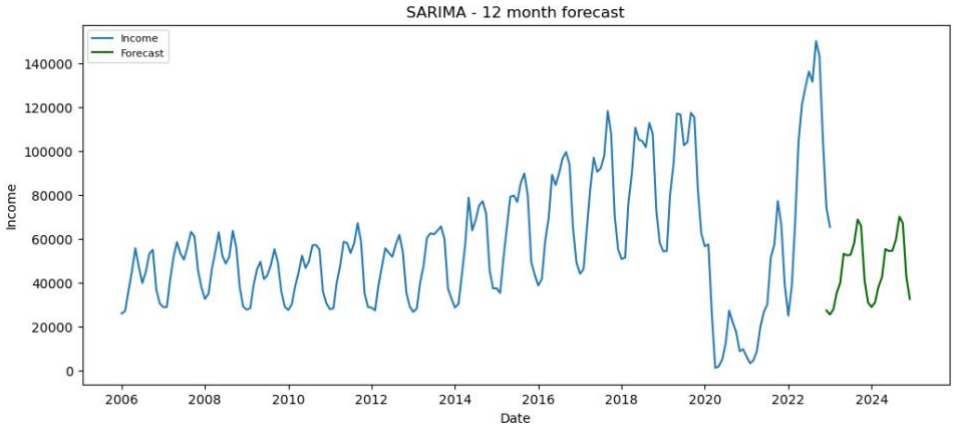

Fig 8. *24 month forecast for 2023 and 2024 using the optimal SARIMA model which minimized AIC.*

Prophet:

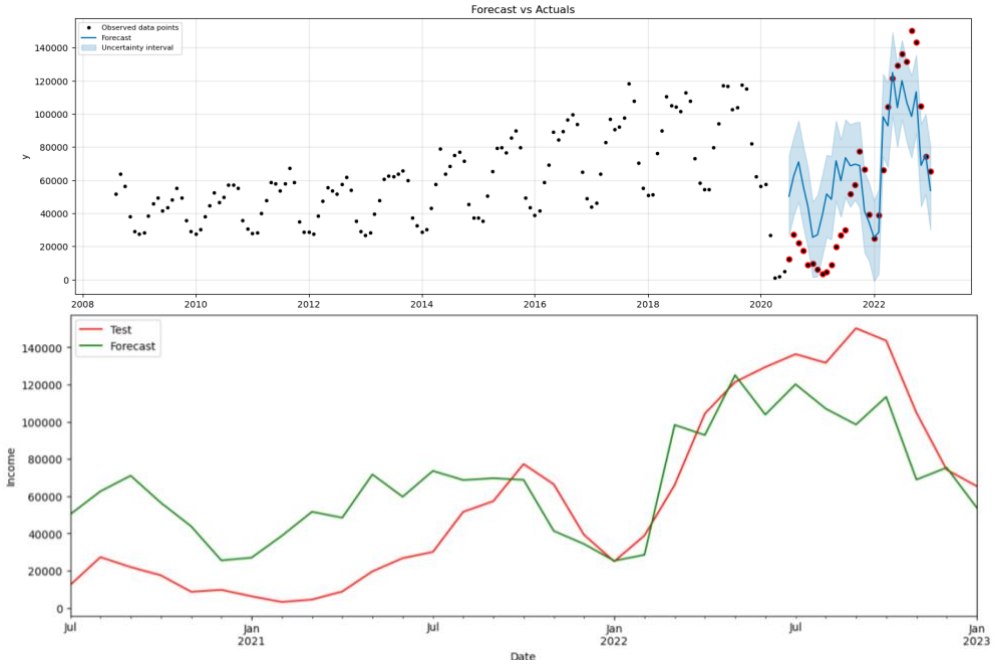

Fig 9. ***(Top)*** *The predicted test values from the raw prophet model with only the holiday parameter entered. The red dots represent the actual data points while the blue line represents the predicted values with confidence intervals.* ***(Bottom)*** *Graphs the predicted values versus the test values for the prophet model.*

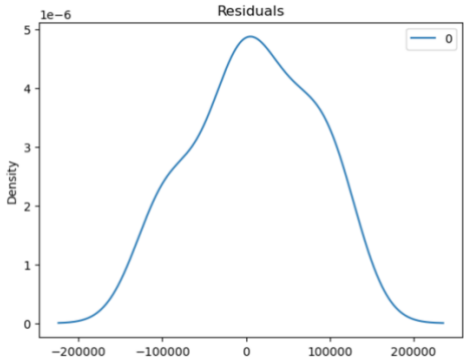

Fig. 10 *Residual density plot for raw prophet model vs test set*

From the above results we were able to obtain a MAPE which was 42%. Although this was a slight improvement from the ARIMA model, we know that we can achieve better results with some hyperparameter tuning. Results a grid search function showed that there were a possible of 150 models that could be considered using our selected hyperparameter. The iterative function suggested that change_point_prior be set to 0.3, holidays_prior_scale be 0.5 and n_changepoints be 100. With this final model that had holidays, tuned hyperparameters and seasonality period set to monthly we carried out cross validation and observed a MAPE of 21%. This is much better than the raw prophet and SARIMA, and even better the model was more accurate at predicting 365 days (12 months) with an average MAPE of 16% shown below in figure 11.

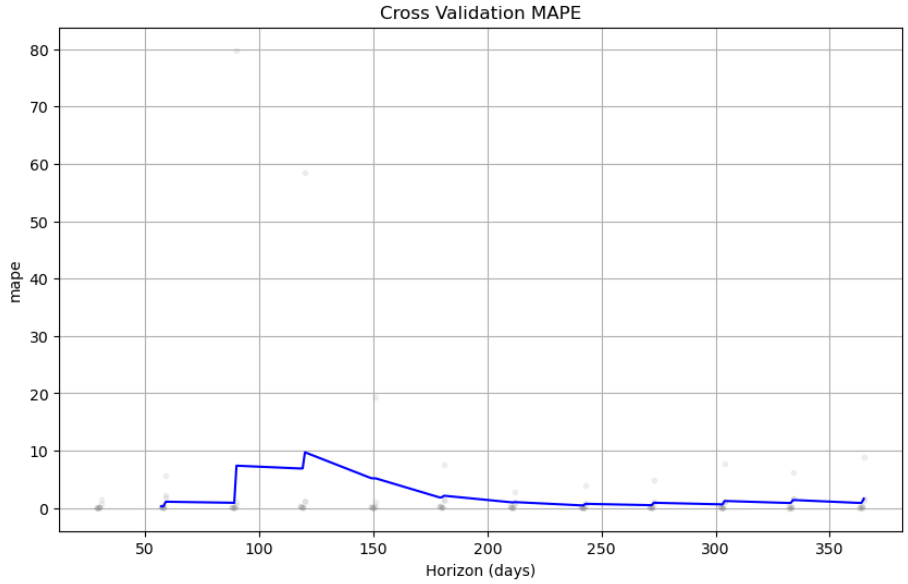

Fig 11. *MAPE values with Prophets cross validation function. As we forecast out to 365 days the model scores better than in the short term.*

We can now fit our prophet model to the time series and forecast into the future. We forecast an extra 11 months to get the 2023 predicted revenue and 23 months to get 2024. These results are plotted below in figures 12 and 13.

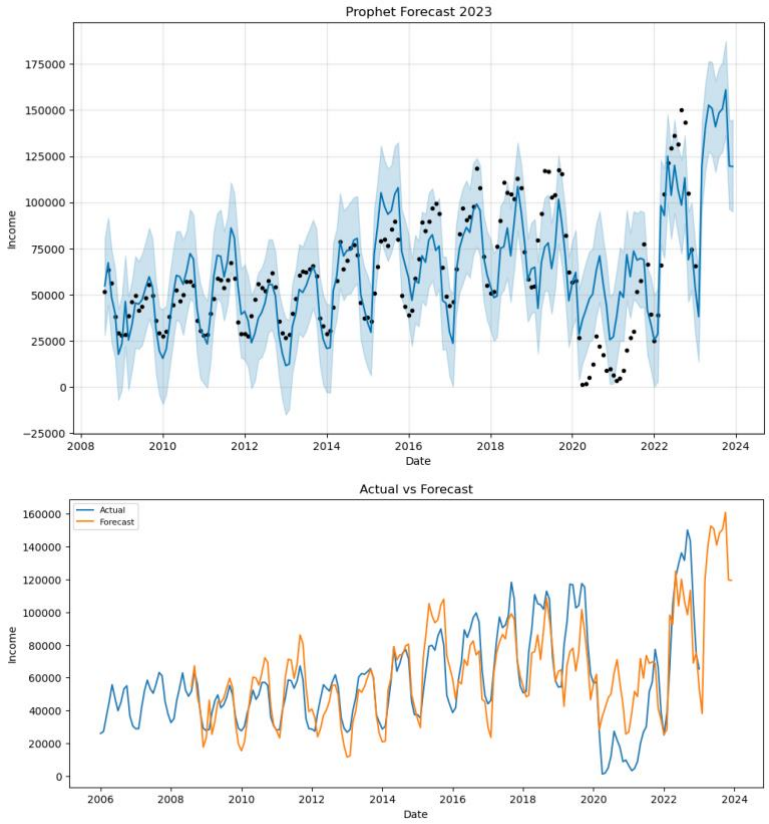

Fig 12. *Above shows the prophet model forecast for 2023. Where each black dot is a datapoint from the actual timeseries and the blue line is our model. Below is a clearer representation of our model (orange line) plotted against the actual data (blue line).*

Above we see the model in orange plotted against the time series. It appears to be a good fit without evidence of overfitting and the effect of covid has a reduced effect.

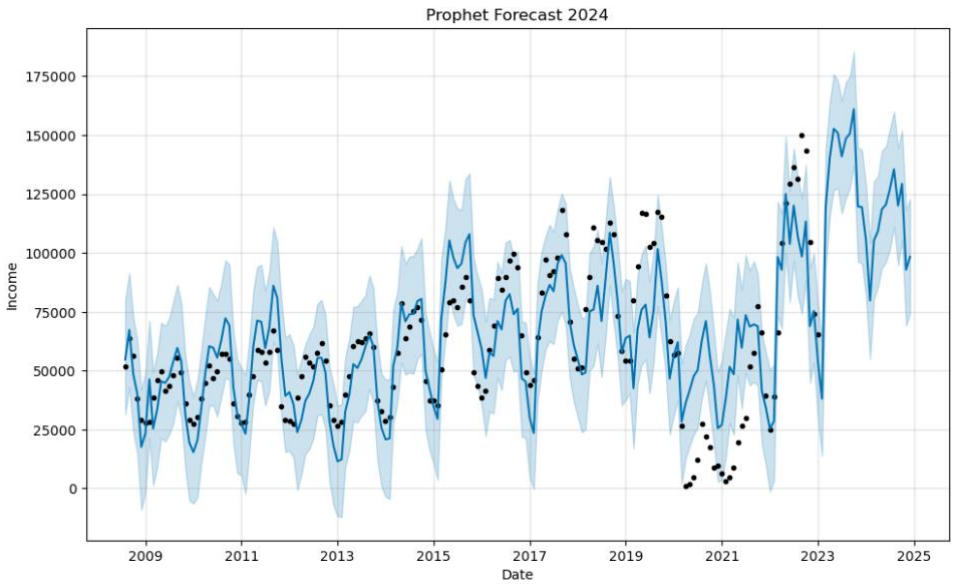

Fig 13. *Prophet forecast for 2024.*

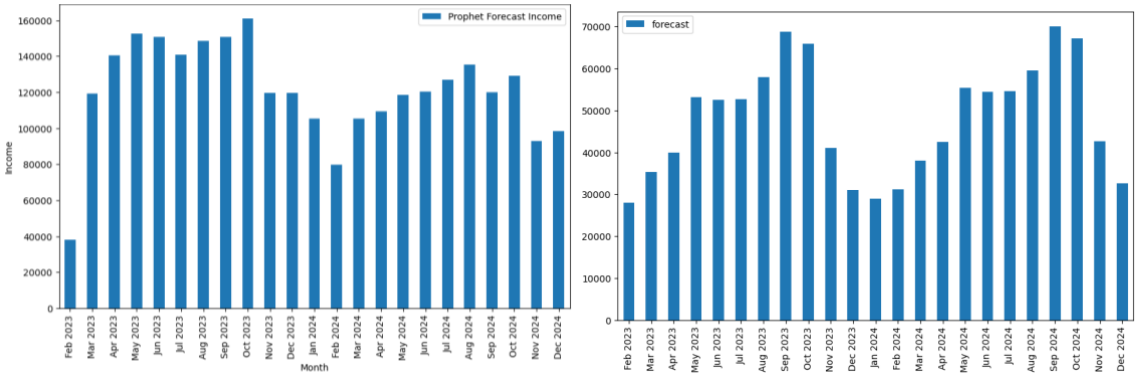

Fig 14. *The forecast values for both the Prophet (left) and ARIMA (right) models. Each month is represented by a bar on these bar charts (note the difference in x-axis). Income is in 000's of euro.*

Figure 14 Shows us the forecasted values for both 2023 and 2024 represented as a bar chart. Both models were able to pick up the seasonal trend which is encouraging. We can see that the prophet model predicted higher values as the weight of the pandemic had less of an effect on the years after the pandemic. The total revenue predicted for year 2023 and 2024 is shown in figure 15. Prophet predicts a total of 1.5 billion in hotel revenue for the city of Lisbon in 2023 and 1.34 billion for 2024. While our ARIMA model predicts 52 million in 2023 and 54 million in 2024. Their differences are also shown below.

Prophet 2023	Arima 2023	Diff.	Prophet 2024	Arima 2024	Diff.
€1,495,469,000	€520,928,000	€974,541,000	€1,341,988,000	€544,481,000	€797,507,000

Fig. 15 *Table showing the sum of forecast values for each model in both 2023 and 2024. The difference is also shown between the model's revenue predictions.*

INE Value (Feb 23)	Prophet Value	% Error	ARIMA Value	% Error
72,640	53,781	25%	27,946	61%

Fig. 16 *Table showing the official Hotel Revenue as reported by INE for February 2023. We have shown both our models forecast for the same month and calculated a percentage error. Revenues are in 000's of euro.*

Discussion:

In the SARIMA Model, it was observed that covid affected the forecasting dramatically. We tried different methods to get around the problem of covid lockdowns. Initially we had deleted the periods during lockdowns and tried mean imputation, but we were unsatisfied with the results and our models were very inaccurate. We eventually decided to leave the covid data in the series and build an ARIMA model. We used an automated function to search for the optimal ARIMA parameters that would yield the lowest AIC. Even with this it was still not enough to limit the effects of Covid on our forecasts. The MAPE value of 46% is a poor result.

For this reason, different models were tested. The Prophet model was preferred, considering that it would be effective for seasonal data in a time series. The holiday parameter in this model removed the effect of the covid lockdowns on the data. We observed a much better MAPE value with the tuning of hyperparameters and cross validation which yielded a MAPE of 16% for yearly forecasts (Fig. 11).

Figure 14 plots the results of our forecast in a bar chart which allows us to visualize the difference between models. The table in figure 15 gives a clear value for the years prediction and even compares the two models. We can see form this that the Prophet model had a much higher prediction than the ARIMA, we strongly believe that this is due to the Prophet models' ability to sidestep the lockdown effects.

What was unique to our project was the fact we had data up to Jan 2023. This meant that when INE released the actual data for Hotel Revenue in Lisbon for February 2023 we could compare our model's accuracy to the official values reported by INE. Figure 16 shows this in tabular form. Here we see that the Prophet model performed much better than the ARIMA. The ARIMA value was extremely low at 27,946 and this was undoubtedly due to the covid impact. This error of 61% was much higher than the Prophets 25% and bearing in mind the prophet models accuracy will only increase at longer horizons so it would be interesting to return to this prediction in the coming months and year.

No model is ever perfect, and with the prophet model there were some limitations. For example, prophet is optimized for daily data and our data was monthly. This is especially evident in the cross-validation function which was difficult to interpret with monthly data. The obvious limitation of the SARIMA model was its inability to deal with extreme events such as covid.

Conclusion:

Taken the above points made in the discussion of the results, it is conceivable to believe that the prophet model is a good model to predict Hotel Industry revenue. The lower MAPE and the better accuracy it showed when compared to ARIMA to the real world INE values for February back up this claim.

The prediction based on the above prophet model suggest that there will be an increase in hotel industry revenue for Lisbon in the next couple of years. This is no surprise judging by the sentiment that Lisbon is new tourist hot spot. The findings above should allow for informed decisions on what areas Lisbon should direct investment such as public transport or tourist attraction budgets. Moreover, it can provide investors with insight into the Hotel Industry here in Lisbon and allow them to make more informed decisions.

References:

1. Chang, X., Gao, M., Wang, Y. & Hou, X. 2012, "Seasonal autoregressive integrated moving average model for precipitation time series", Journal of Mathematics and Statistics, vol. 8, no. 4, pp. 500-505.

2. Feng, T., Zheng, Z., Xu, J., Liu, M., Li, M., Jia, H., & Yu, X. 2022, “The comparative analysis of SARIMA, Facebook Prophet, and LSTM for road traffic injury prediction in Northeast China”. Frontiers in public health, *10*, 946563.

3. Santos, N. & Oliveira Moreira, C. 2021, "Uncertainty and expectations in Portugal's tourism activities. Impacts of COVID-19", Research in Globalization, vol. 3

4. C. J. Costa and J. T. Aparicio. 2020, "POST-DS: A Methodology to Boost Data Science," 15th Iberian Conference on Information Systems and Technologies (CISTI), Seville, Spain, pp. 1-6, doi: 10.23919/CISTI49556.2020.9140932