Demand Prediction for Shared Mobility Services using Time Series Modelling

Rudi Camilleri[1] and Jeremy Debattista[2] [0000-0002-5592-8936]

[1] University of Malta, rudicamilleri@gmail.com
[2] jerdebattista@gmail.com

Abstract. The main objective of this paper is to analyse and investigate the possibilities of optimising shared mobility using historical data by predicting the total number of generated requests per hour. The study first investigates where and how pickup requests were made. Subsequently, the data is examined for any time-series patterns, primarily trend, seasonality and cyclic. Different types of parametric techniques like Holt-Winter, ARIMA and Facebook Prophet were used to compare whether seasonality and other characteristics have any effect on the performance of the model. From the results obtained, it can be implied that exogenous or independent data such as temperature and public holidays do not affect the predictive model. Metric performance results conclude that amongst all models, the accuracy of Holt-Winter's outperforms other models with an overall MAE of 8.039 and RMSE of 11.159.

1 Introduction and Background

Shared mobility is an evolutionary force of modern transportation. Options include bike sharing, ride-hailing, and carpooling. Traditionally, travellers use public transport to go to work, or private vehicle for other commutes. Unfortunately, people's desire for private cars has led to several environmental side effects such as excessive delays and congestions. Malta, as one of the smallest countries in the world, it is also one of the most densely populated countries with around 1,562 people per square kilometre. As population density increases, so does traffic congestion. The mean amount of car trips per week in Europe varies between 2.4 and 2.9, whereas in Malta it is 3.2 trips per driver. Facing this barrier, the concept of shared mobility has emerged as the key model to reduce traffic congestions, lower emission rates, and boost the accessibility for travellers. The operation of ride-sharing management can be improved by making sense out of existing data in a statistical manner. The purpose of predicting future demand using different means of data is to identify patterns and learn from the information you already have. Such data could be geographical coordinates and geosocial data.

The motivation for this paper is to identify and develop different forecasting models which can predict hourly ride-sharing demand in Malta. The use of different statistical time series techniques, such as linear methods and exponential smoothing can help in optimising shared mobility. Models will use past data generated by any IoT devices and other variables that might have a relationship with the demand. External variables such as weather information which can influence the demand can be considered as an asset for the models' accuracy. With patterns like trends and seasonalities, the outcome should be capable of providing a close idea of what the near future holds. Since a common event can repeat in various frequencies like hourly, daily, or yearly, characteristics like seasonality can distinguish patterns at a specific time.

Time series modelling is a type of time-based information that can be utilised to visualise different mediums. Its study is more popular than ever before as computers' computation power continues to improve. Nonetheless, the analysis of time series has yet to reach its best time, as traditional statistical methods still lead it. When forecasting, historical data is explored, analysed, and structured to anticipate the future. However, there is no general rule in what type of predictive technology to use, as what might work in economics does not mean it will work within the transportation industry. In this case, the primary algorithms used are the Autoregressive Integrated Moving Average (ARIMA) [1], Facebook Prophet [13] and Holt-Linear [12].

A time series mainly differ from one another due to its trend, cyclical behaviour, seasonal fluctuations, and irregular changes. Another precept in time series forecasting is stationarity, and for a series to be considered stationary, its statistical properties, mean, variance and covariance, should be constant throughout. A simple method to convert a series to become stationary is by using differentiation. With the use of a unit root test, one can identify whether stationarity is achieved. To identify whether values are randomly placed or not, a lag plot is used. The correlation between data points depends on the pattern shown by the graph. The same relationship can be identified using both the Autocorrelation (ACF) and Partial Autocorrelation Function (PACF). Whenever a series is substantially autocorrelated, observed data can be crucial in predicting future values.

Popular forecasting methods for data with seasonal patterns and trends are the Triple Exponential Smoothing, Seasonal Autoregressive Integrated Moving Average (SARIMA) and Prophet. Other models which will be tested throughout the project are the Moving Average, Autoregressive, ARMA, and SARIMAX. The latter model makes use of exogenous variables to help in optimising the predictive accuracy. Similar to ARIMA, the SARIMAX is a generalized form of the ARMA model. Prophet, a predictive model developed by the core data science team at Facebook, makes use of a forecast evaluation framework called 'Analyst in the Loop'. The algorithm is based on a Generalized Additive Model (GAM) and like the Auto-Regression model, non-linear trends are fit with yearly, weekly, and daily seasonality. Furthermore, despite the use of exogenous variables, the model also accepts holiday and event dates for data fitting.

Different performance metric measurements are used to understand how accurate a predictive model is. Popular key performance indicators in this field are the Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE). According to Armstrong [14], calculating error measurement is not easily accomplished, as there is no one-size-fits-all method. One metric may behave well in one scenario but fails in another. Therefore, the best indicator can only be established by experimentation. The second section discusses the challenges, objectives, and hypothesis. Section 3 explains the overall framework of the study, while the 4th section is where the models are developed. In section 5, results are compared with the observed data.

2 Challenges and Objectives

The main aim of this project is to optimise shared mobility using multiple predictive techniques and historical data as provided. The challenges tackled are as follows:

C1: Developing and identifying the right time series techniques

O1. A time-series is developed and decomposed to identify certain parameters, such as trend, seasonality, and residuals.

C2: Understanding and modelling car-sharing data

O1. Data coordinates are plotted on a geographical map. The map is clustered in groups for better analysis and vehicle distribution in the future.

O2. The predictive models are trained using the data provided from the local operators, and other exogenous variables are also included as they might improve the accuracy of the same models.

C3. Determine the most accurate models

O1. Generate a performance metric report for all models.

O2. Compare the MAE, MSE and RMSE metrics to determine the best model.

Modern algorithms, like the Facebook Prophet, SARIMA, and Holt-Winters [12] have the capability of training models using seasonality. Moreover, such models have other characteristics which as a model, make them unique from one another. Therefore, the hypothesis for this paper is:

Time-series models with the seasonality predictor provides better predictions for shared mobility demand than those without the seasonality predictor.

3 Related Work

The study of time series data is becoming extremely significant in all sectors, with tech innovation and big data in for example digital health, internet of things and smart cities. Over the years, a lot of research and studies have been done by researchers to improve forecasting accuracy and develop better models for time series analysis.

Wang [2] compared a Convolutional Neural Network (CNN) with a Long Short-Term Memory (LSTM) to predict a Chinese ride-hailing service from 7 million records of trips. The author divided the Chinese city, Chengdu, into smaller regions and later included context information like weather data to improve the prediction performance. In another research [3] tackling short term passenger demand forecasting, the authors implied that spatial, temporal, and exogenous dependencies make the forecasting models more challenging. Similar predictive problems proposed using an ARIMA model [4] [5], local regression model [6], neural network-based algorithm [10], and

Bayesian linear model approach [7]. As seen in [8], hybrid models are used as a combination of multiple predictive techniques, to improve the prediction's accuracy. Another study [9] investigates the impact of holidays and events compared with normal days. V. Shah [11] concluded that the model for a quantitative method with external parameters is SARIMA while without external parameters than it is Holt-Winters. Despite SARIMA performing well, its drawback is the computational power to fit the dataset.

The analysis of time series data is becoming extremely significant in all sectors, especially with smarter cities. This study suggests that by integrating together various techniques used in other papers, the predicted results will be more accurate and reliable.

4 Design

In this study, as seen in the framework below, the steps taken are data collection and pre-processing, data visualisation, applying stationarity to a timeseries and finding its optimal parameter combinations, train and build the model, and predict and test the result.

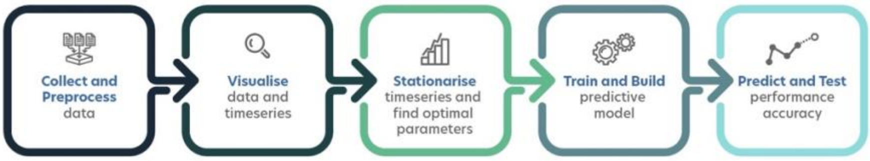
Collect and Preprocess data
Visualise data and timeseries
Stationarise timeseries and find optimal parameters
Train and Build predictive model
Predict and Test performance accuracy

Fig. 1. Workflow to extract and make sense out of car-sharing data

In the first step, the anonymised data granted by a local Maltese ride-sharing operator was processed in a way that is easily managed by Python libraries. The dataset consists of 92,082 total number of requests throughout the period of December 2019 till the beginning of March 2020. Table 1 briefly explains the 11 variables present in the dataset.

Table 1. Ride-sharing dataset

Variable Name	Description
RIDER_ID	An automatically generated integer type variable used to distinguish between users. Ranges between 28 and 38,135.
RIDE_ID	Automatically generated integer variable for every request. Ranges between 73402 and 177191.
REQUEST_DATETIME	Date and time for the request taking place.
REQUEST_STATUS	Categorical variable explaining whether a request was Completed or Cancelled.
PICKUP_DATETIME	Date and time for the pickup taking place.

ORIGIN_LAT	Pickup latitude coordinate.
ORIGIN_LONG	Pickup longitude coordinate.
DROPOFF_DATETIME	Date and time for the drop off taking place.
DESTINATION_LAT	Dropoff latitude coordinate.
DESTINATION_LONG	Dropoff longitude coordinate.
DISTANCE_KM	Trip distance in total.

A secondary dataset which will be used is a weather time series consisting of data points throughout the same period as of the ridesharing one, between December and March 2020. This data was obtained from 'Maltese Island Weather' and it is based on an hourly basis. The present variables are Mean Temp (°C), Rel. Humidity (%), Mean Wind Speed (km/h), Wind Gust Speed (km/h), and Rainfall (mm). This will be investigated to see whether weather data has any effects on shared mobility.

Undeniably, the first process is to read the raw dataset in question. Considering that this project will only cater for successful requests, a total of 3,875 cancelled requests were removed as for this study the end goal is to predict successful trips. Furthermore, variables containing date and time information, are converted from an object type to a DateTime type. This will facilitate the process later, as such information will be used as a time series. In addition to this, unwanted data like the 'RIDER_ID' and 'RIDE_ID' were removed as well. Figure 2 shows the relationship between the number of daily requests and time.

For the sake of accuracy, irrelevant data such as outliers and cancelled requests are removed. Outliers are determined by plotting a box plot and calculating the percentile values for the average speed, waiting time, trip time and trip distance. Since percentiles are remarkably significant in identifying outliers, the quantile function was used, and the data points were distributed in quantiles relative to 100. Furthermore, removing these outliers ensured that the data is cleaned out from erroneous points. In this case, the flawed value was always the 100th percentile, with 278 minutes in waiting time, 661 minutes of driving, and 21.65 distance km.

Secondly, it is time to visualise both geographically and linearly the data to distinguish the characteristics of such time series. A heatmap suggested that the busiest areas are mainly northern touristic areas like Sliema Valletta, and Luqa next to the airport. Moreover, by plotting a correlation matrix and a line chart, it was found that Friday and Saturday evenings are the most hectic. At the same time, weekdays from Monday to Friday showed an identical pattern.

In data clustering, the elbow method recommends two groups, and consequently, the separation is implemented using the K-Means algorithm. 71.27% of the cleaned data is grouped with the first cluster, while the rest form part of the second cluster.

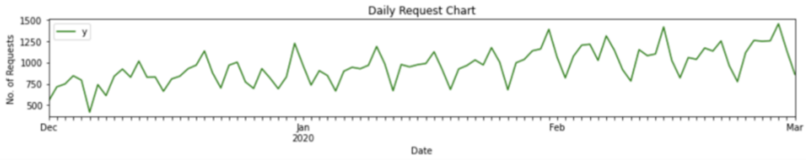

Fig. 2. Observed data plot

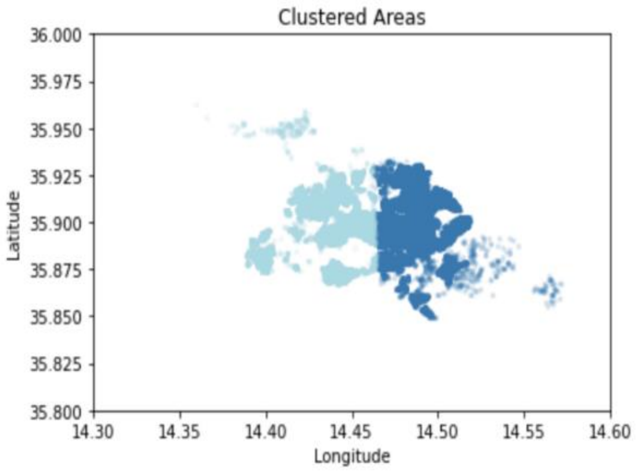

Fig. 3. Clustered Area

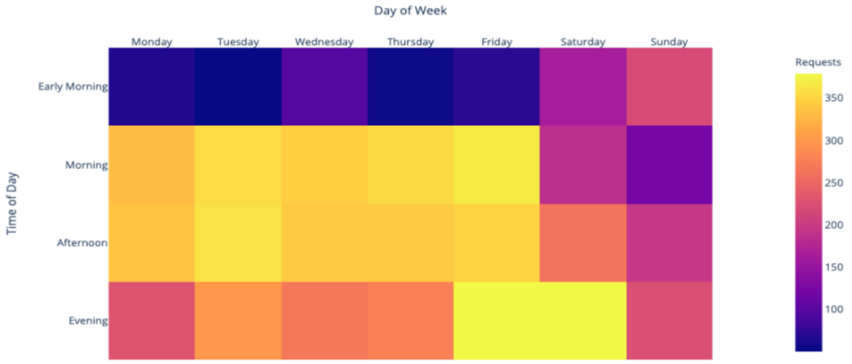

Fig. 4. Correlation heatmap

For a better user experience in forecasting, an intuitive interface was developed using the Tkinter library. Using Python's GUI toolkit, Tkinter, an intuitive interface is developed for the end-user to forecast results. First, the user imports the data to be processed and cleaned. Secondly, since this is a proof of concept and the models are not yet perfect, it was decided that the user will have the option to choose his preferred predictive model. After the model is trained on the data, the user must select the starting date and the number of hours to be able to generate a report. Eventually, the result is exported as a generated table in a pdf document.

5 Implementation

This section will explain how various models are prepared for data forecasting. After all the necessary processing is done, the data is split between a training set and testing set. When assessing a forecasting model, it is important that the predicted data is only compared with the data from the testing set. In machine learning, this is known as the out of sample data. In this case, the training set has two months of data while the testing has one month.

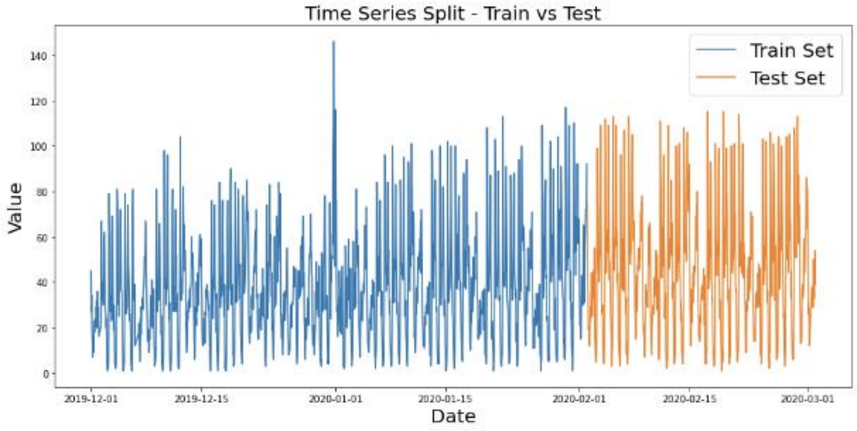

Fig. 5. Train and Test Split

Eventually, the time series is decomposed for its level, trend, seasonality, and noise using the seasonal_decompose function. The abscissa of the lower peak from the seasonal plot, suggests that the series has daily seasonality. Moreover, the lag plot below shows a moderate auto-correlation pattern.

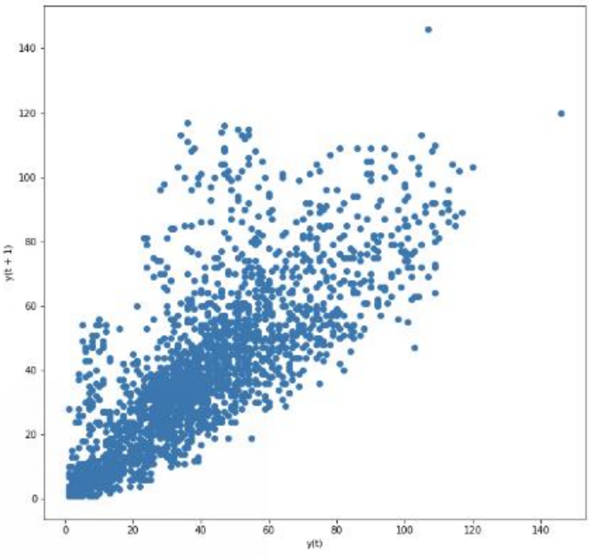

Fig. 6. Lag Plot

Various model combinations are tested using the auto_arima function from the pmdarima class. Such models include the ARIMA, SARIMA and SARIMAX. The combination with the least Akaike (AIC) value is the best from all other option. Unlike the ARIMA, the Prophet model takes into consideration the seasonal effects caused by peaks on public holidays and events. On the other hand, the procedure to fit the Prophet model is similar to the ARIMA. In addition, a triple exponential smoothing function, holt-winters, was trained with a seasonal period of 168.

6 Evaluation and Results

The scope of this research is to assess different time series modelling techniques and predict ride hailing requests by the hour. The plan is to make use of exogenous data that could help in improving the forecasting accuracy. When all is said and done, the operators will be capable of getting rid of unwanted vehicles from the streets by only being at the right place at the right time. Ahead of any predictive techniques, the data is cleaned from irrelevant information, from which a total of 4,754 (5.44%) observations were removed for having a cancelled request status or for being classified as outliers. For the outcome, the dataset incorporates a total of 87,328 data points through a 3-month period.

Since one of the main objectives of this research was to investigate and identify a good forecasting model, multiple forecasting techniques were developed. Moreover, exogenous values were used to investigate the possibility of improving the prediction accuracy of the model currently being used. Performance metric calculations were done on various timestamps, varying between days, weekends, weeks, and months.

Taking in consideration the MAE and RMSE for all models, Holt-Winter seems to give a better result for the shorter term. In Holt's, the non-weighted estimate, MAE, returns a mean of less than 9 while the maximum weighted estimate for the first week is 11.94. Moreover, both Prophet's estimates reach almost 17 and 22, respectively. The following Friday, the model misplaced by 25, few more than the 15 predicted by Holt. Furthermore, Holt has the best score for the 2nd week, with a total difference of 304 requests in a week from the 7,418 total observed requests. In this case, the Prophet misplaced the actual week value by 417, while other models varied between 1085 and 3255 differences in a week. From all the calculations done, it can be concluded that Holt-Winters managed to achieve the best performance score.

7 Conclusions and Future Work

This study was put together to figure out whether seasonal models provide better mobility demand prediction than other models. Including the Maltese ride-sharing data, weather data was also obtained from a local weather station. To understand the time series better the data was decomposed into various components of a time series, mainly, level, trend, and seasonality. It was noted that during the week, most pickup requests were recorded during rush hours. As expected, the quietest period happened to occur during the night and early hours in the morning. Moreover, the highest number of requests were recorded on Fridays and Saturdays, with a daily average of 360.08 and 361.46 requests, respectively. During data processing, data was converted in different data types, split between training and testing set, and unnecessary outliers were removed. Geo-location data was divided in clusters for future reference when developing the report generator.

Various predictive models, such as Holt-Linear, SARIMA, and Facebook Prophet, were trained and when possible, used the weather data mentioned earlier, to try and help in improving the accuracy of the same models. With the help of a performance metric report, it was concluded that the weather did not have any effect on ride-sharing demand, as there was no improvement in the accuracy. It was established that Holt-Winter surpassed all other predictive techniques used in this study, for both the short term and longer term. In certain timeframe scenarios, FB Prophet, was the best predictor when no exogenous data is used. Therefore, with seasonal models outperforming other traditional non-seasonal ones, the hypothesis was validated.

7.1 Future Work

Investigating more complex techniques such as Neural Networks might result in better performance. Additionally, social media can be used as an extra exogenous variable, to help in optimising demand prediction for shared mobility services. Finally, the report generator can be improved in many ways such as embracing the use of real-time data streaming. Secondly, the generator can process all the models internally, however by using different performance metrics, it will output only the report generated by the best fitted model.

References

1. G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, Time series analysis: forecasting and control, vol. 734. John Wiley & Sons, 2011.
2. C. Wang, Y. Hou, and M. Barth, “Data-Driven Multi-step Demand Prediction for Ride-Hailing Services Using Convolutional Neural Network,” Adv. Intell. Syst. Comput., 2020.
3. J. Ke, H. Zheng, H. Yang, and X. (Michael) Chen, “Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach,” Transp. Res. Part C Emerg. Technol., vol. 85, pp. 591–608, 2017.
4. M. Khashei, M. Bijari, and G. A. R. Ardali, “Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs),” Comput. Ind. Eng., 2012.
5. N. Zhang, Y. Zhang, and H. Lu, “Seasonal autoregressive integrated moving average and support vector machine models: prediction of short-term traffic flow on freeways,” Transp. Res. Rec., vol. 2215, no. 1, pp. 85–92, 2011.
6. C. Antoniou, H. N. Koutsopoulos, and G. Yannis, “Dynamic data-driven local traffic state estimation and prediction,” Transp. Res. Part C Emerg. Technol., vol. 34, pp. 89–107, 2013.
7. X. Fei, C.-C. Lu, and K. Liu, “A bayesian dynamic linear model approach for real-time short-term freeway travel time prediction,” Transp. Res. Part C Emerg. Technol., vol. 19, no. 6, pp. 1306–1318, 2011.
8. M. Khashei, M. Bijari, and G. A. R. Ardali, “Hybridization of autoregressive integrated moving average (ARIMA) with probabilistic neural networks (PNNs),” Comput. Ind. Eng., vol. 63, no. 1, pp. 37–45, 2012.
9. M. Cools, “Investigating the Variability in Daily Traffic,” no. X, pp. 1–22.
10. K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang, “Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and Levenberg--Marquardt algorithm,” IEEE Trans. Intell. Transp. Syst., vol. 13, no. 2, pp. 644–654, 2011.
11. V. Shah, “A Comparative Study of Univariate Time-series Methods for Sales Forecasting,” 2020.
12. P. S. Kalekar and others, “Time series forecasting using holt-winters exponential smoothing,” Kanwal Rekhi Sch. Inf. Technol., vol. 4329008, no. 13, pp. 1–13, 2004
13. S. J. Taylor and B. Letham, “Forecasting at scale,” Am. Stat., vol. 72, no. 1, pp. 37–45, 2018.
14. J. S. Armstrong and R. Fildes, “Correspondence on the selection of error measures for comparisons among forecasting methods,” J. Forecast., vol. 14, no. 1, pp. 67–71, 1995.