Forecasting tourist arrivals with Machine Learning

Shaikh Aazar

17th December, 2021

Abstract

Forecasting is a concept of attempting to predict future development of a process or phenomenon using data previously available and appropriate mathematical techniques and models. Substantial short-term demand fluctuations are common in the tourism industry. Therefore, tourism companies such as accommodation, transportation, catering, and leisure facilities have a vital interest in precise forecasts of the number of customers.

1. Introduction

1.1 Background

Tourism defined by United Nations World Tourism Organization (UNWTO) as

> ***Tourism*** *comprises the activities of persons travelling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business and other purposes.*
>
> ***Tourists*** *are visitors who stay at least one night in a collective or private accommodation in a place visited.*
>
> *The* ***same-day visitor*** *is a visitor who does not spend the night in a collective or private accommodation in the place visited. This includes cruise passengers who debark in a country but spend their nights in a board ship.*
>
> ***Tourism expenditure*** *is the total consumption expenditure made by a visitor on behalf of a visitor for and during his or her trip and stay at a destination*
>
> *The* ***tourism industries*** *designate the set of enterprises, establishments and other organizations one of whose principal activities is to provide goods and/or services to tourists*

Tourism is one of the world's largest industries. It contributes significantly to economic growth, contributing 10% of the world's GDP, the most comprehensive

measure of the total value of the goods and services the world's economies produce.

Thus, forecasting tourist volume is becoming increasingly important for predicting future economic development. Tourism demand forecasting may provide basic information for subsequent planning and policy making.

Methods used in tourism modelling and forecasting fall into four groups:

1. Time series Models
2. Econometrics models
3. Artificial Intelligence Techniques
4. Qualitative methods (expert-opinion techniques)

Time series models are the simplest and least costly and expert-opinion methods are useful when data are unavailable. In addition to simple tourist data announced by the State Statistics Bureau, Internet search queries, which reflect the behaviour and intentions of tourists, have increasingly been used in tourism forecasting models.

1.2 Purpose

- The tourism product is perishable in nature i.e., only active for a brief period of time for generating revenue and inactive other (most) times.
- Its demand is vulnerable to many external factors (consider Covid lockdowns for example).

These two features make the "forecasting of tourism flows" essential.

The purpose of this service would be to get estimate of future demand at destination(s) level which would then help in managing and planning tourism development and the necessary investment.

1.3 Objective/Problem Statement

- To build a Business to Business (B2B) Tourism Forecasting model using methods most suitable for regional tourism organizations and private companies.
- Model should implement trending Internet Search Queries to help keep track of latest trends which may provide more accurate and predictions.

2. Assessment

2.1) Customer Needs

Customer for this service/product will usually be owners/managers of businesses in tourism industries (defined above)

- Accuracy: estimates provided by the model should conform to the actual event being forecast with an acceptable confidence interval.
- Cost: The model is aimed for regional tourism organizations or private companies at lower levels, so cost should be economical
- Robustness: the mathematical strategies used in the model should be as such that it is not sensitive to the random extreme values ('outliers') in the training data set.
- Structure should be properly specified

2.2) Business Needs

- Forecasting for better investment opportunities
- Analysing trends and doing business accordingly can be very profitable.

3. Target Specification

1. **Access** in terms of development and maintenance of transport which provides the link to the tourist destination as well as the tourist attractions at the destination.
2. **Attractions** which motivate and attract tourist to visit the destination and it consist of the man made as well as natural attraction features or cultural events.
3. **Hospitality sector** – comprise of accommodation organization, catering organization and attitudes of community towards tourists and tourism business.
4. **Organization Sector** – comprise of all the operations within the tourism distribution system who determine the movement of travel packages from manufacturers to tourists through intermediaries and support services.

These can be achieved by planning, preparations, managing, and maybe even investing with using proper forecasting.

4. External Searches

4.1) Time Series Models: these are simple to use and most popular, the naïve time series forecasts simply extrapolate past data and can be used to identify

seasonal variation and trends. Time Series forecasts are usually univariate: i.e., they forecast a variable based on past data for that variable

4.2) Econometric Causal Model: attempts to measure cause and effect relationships among variables. Techniques involve standard multivariate regression analysis, discriminant analysis and probit analysis.

4.3) Expert Opinion Methods: Handling qualitative variables with subjective judgements, includes Feedbacks, Surveys etc.

5. Benchmarking

Type	Data	Cost	Simplicity	Users
Time-Series	At least 2 year's data	Low	Simple	Suitable for private suppliers
Econometric-causal models	> 2 years'	Medium	Complex	Useful at regional Level
Expert-opinion Techniques	Little or no data required	Low	Moderate	All users

6. Applicable Patents

The most relevant patents I could find.

- Tourism management system
- Tourism information and reservation system and method

7. Applicable Regulations

- Privacy and Data Protection (tourists)
- Insurance Claim Regulations/Policies
- Patents/Rights on Algorithms developed
- Regulations regarding handling collected data (breach must never occur)

8. Applicable Constraints

- Obtaining datasets for providing accurate results can be challenging

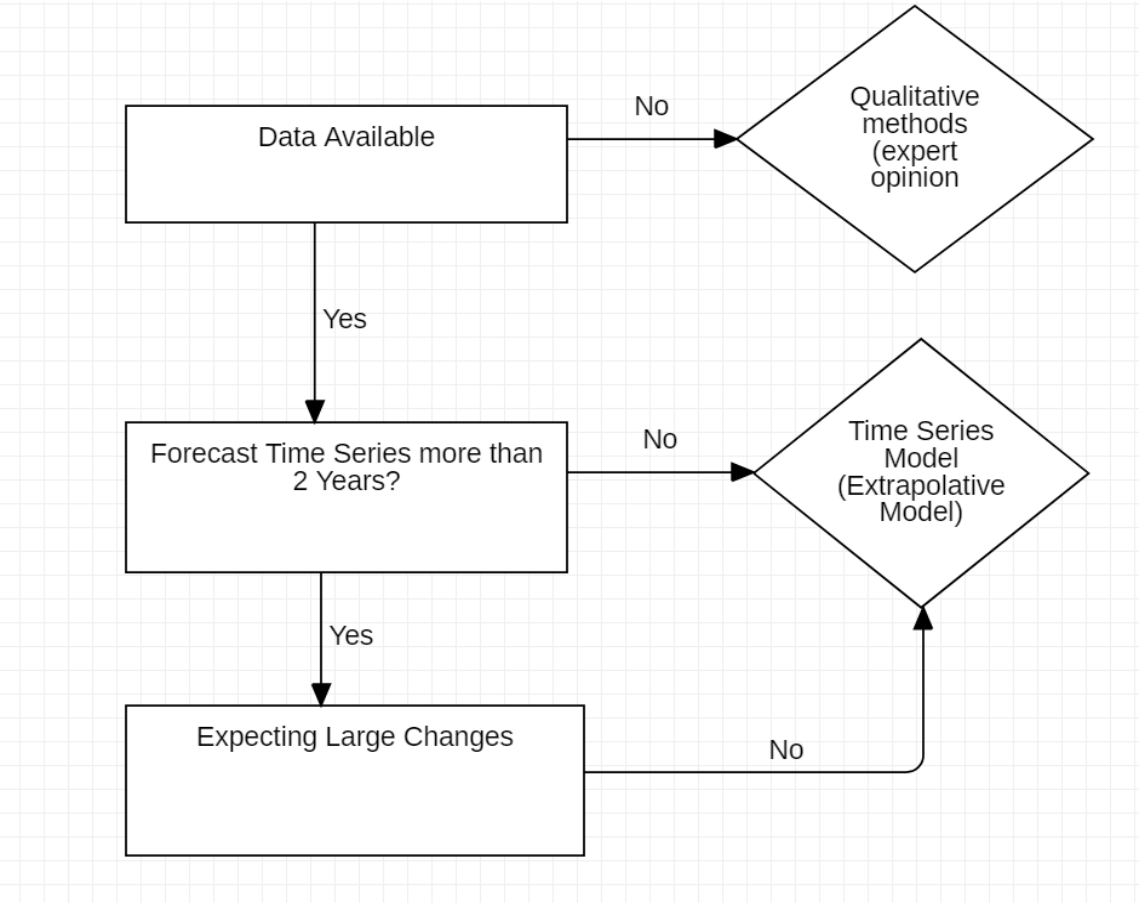
- Acceptance – reliability of software is not a given so establishing the product may be difficult
- Regularly updating data and training the model simultaneously and maintaining accuracy within a threshold confidence interval.

9. Business Opportunity

1. **Tourism –** Directly aimed to enhance tourism, specially at regional level or private companies
2. **Transportation –** Mainly public transport system, specially near tourist attractions.
3. **Businesses which complement tourism.**

10. Concept Generation

From target specification and analysis of different models of tourism forecasting, it can be concluded that for this particular set of requirements, the ideal implementation model to choose would be **Time-Series** model of forecasting.



11.Final Prototype Design

Extrapolative methods (Time-Series) use patterns in the data during the past to project or 'extrapolate' future values, i.e., we need to first collect historical data to train our model (at least 2 year's data for sufficient accuracy/confidence intervals).

1. Collecting preparing and verifying input data
2. Programming the initial model we have chosen
3. Estimating model parameters
4. Determining model's accuracy in the past

   4.1. Mean absolute percentage Error (MAPE)

$$MAPE = \frac{1}{n} * \left(\frac{|e|}{A}\right) \times 100$$

W*here, n = number of periods*

e = forecast error

A_t *= Actual value of the variable being forecast*

t = some time period

Interpretation of MAPE values:

- Less than 10% is highly accurate forecasting
- Between 10-20% is good forecasting
- Between 20-50& is reasonable forecasting
- Greater than 50% is inaccurate forecasting

4.2. Root Mean Square Percentage Error

$$RMPSE = \sqrt{\frac{\left(\frac{e}{A}\right)^2}{n}} * 100$$

Where, n = number of periods

e = forecast error

A_t *= Actual value of the variable being forecast*

t = some time period

5. Publishing Prediction Intervals of forecast – while predicting certain value in future, we are predicting only a likelihood among many other possible values,

so we consider a range of possible values about which we are certain about, this range of values is called ‘prediction interval’.

$$F_{n+h} = \pm z * \sqrt{MSE_H}$$

$$MSE_H = \frac{1}{(n-h)} \times \sum_{t=h+1}^{n} e_i^{h\,2}$$

Where, F_{n+h} = prediction interval h periods after last value in historical time-series data

MSE_h = Mean Squared error of the forecast value h periods after last value in historical time series

z = level of confidence taken from normal distribution
n = number of values in historical time series
h = number of periods after the last value in historical time series
e^h = error for a time period raised to the power h
t = some time period

6. Obtaining the forecast

Tourism Dataset used for sample/basic implementation of algorithms.

```
[2]: other = pd.read_csv("data.csv")
     other.head()
```

[2]:

	x_c_dat1	accommodation	transport1	transport2	transport2_mask	indoor_leisure	indoor_leisure_mask	x_c_d1	x_c_d15
0	39083	44.0	0	3621.0	0	1867.0	0	1	1
1	39084	39.0	0	5222.0	0	2615.0	0	0	1
2	39085	50.0	368	4169.0	0	2241.0	0	0	1
3	39086	35.0	57	4252.0	0	2470.0	0	0	1
4	39087	26.0	248	5059.0	0	2309.0	0	0	1

5 rows × 562 columns

Evaluating parameters for AREMA Modelling

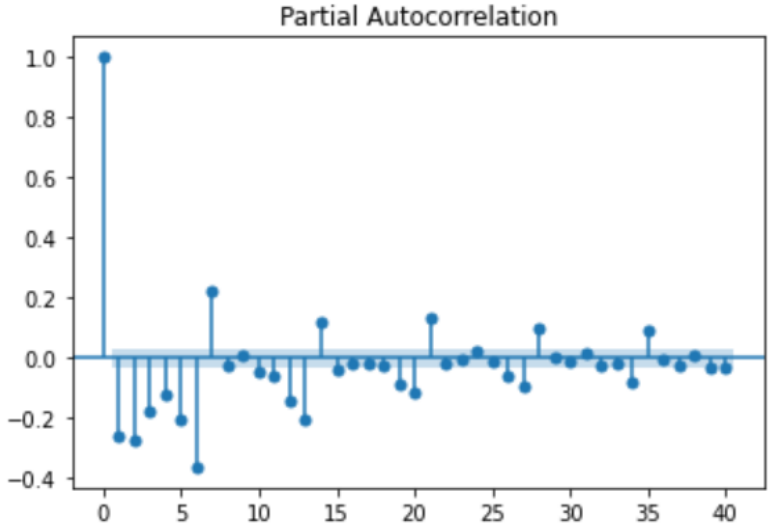
```
: from statsmodels.tsa.stattools import adfuller
```

```
: adfuller(feature1.dropna())       # Calculating p-value for ARIMA parameters
```

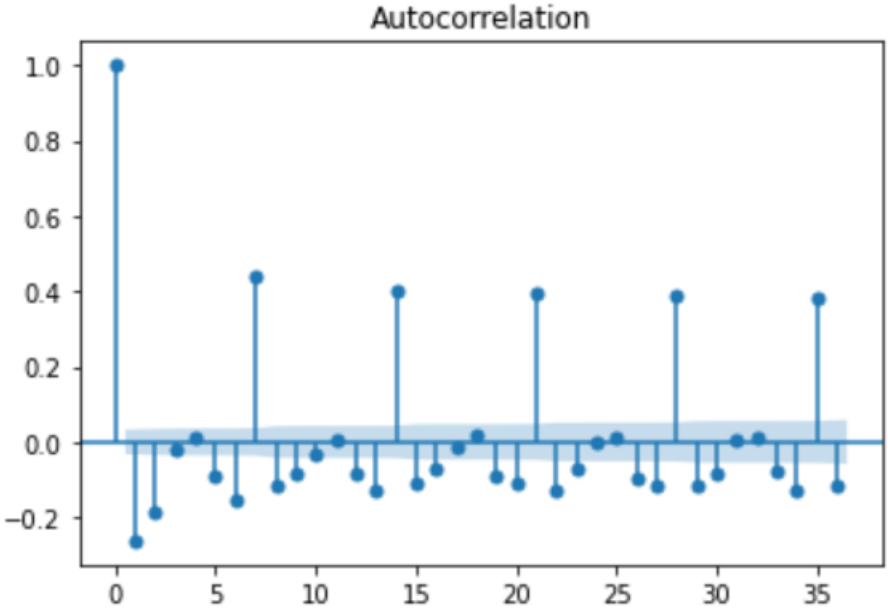
```
: 7.322228471153604e-05
```

```
: # d = 0
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
plot_pacf(feature1.diff().dropna(), lags=40)
```

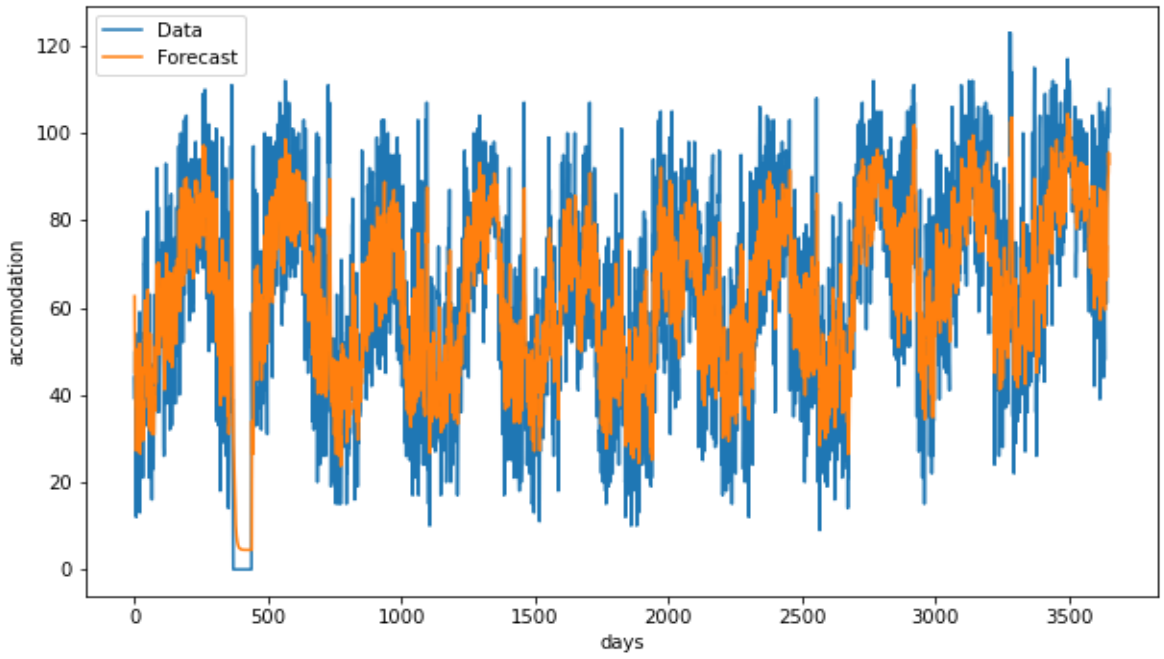
:

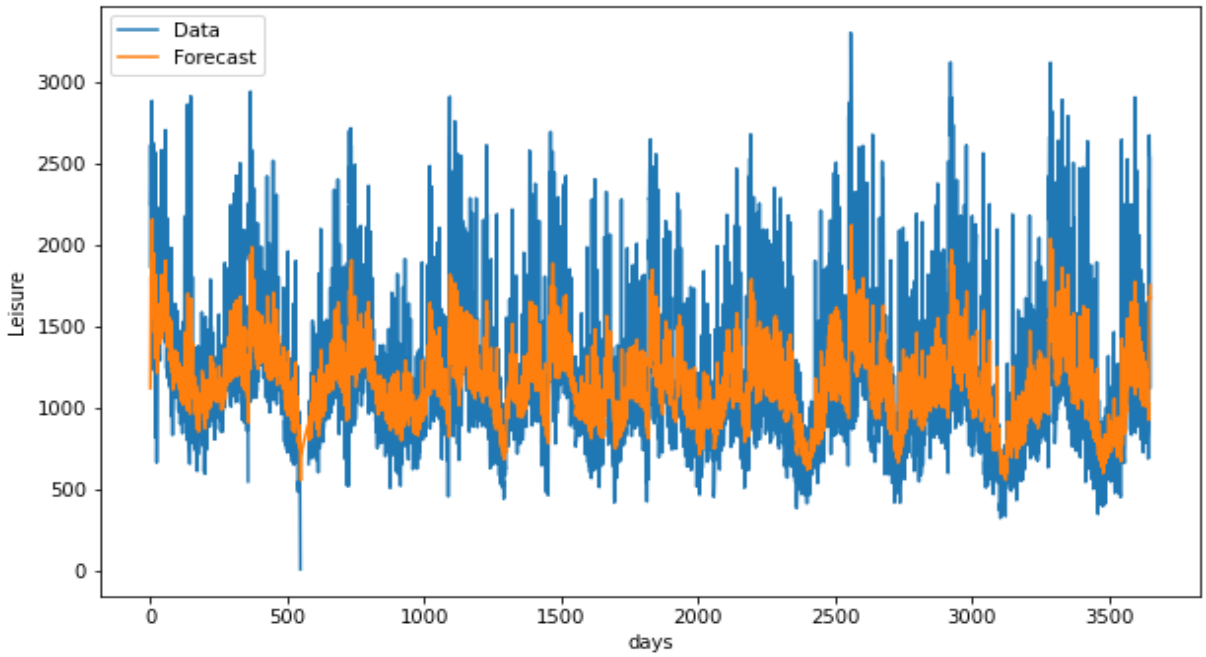



```
# let p = 1
plot_acf(feature1.diff().dropna())
```

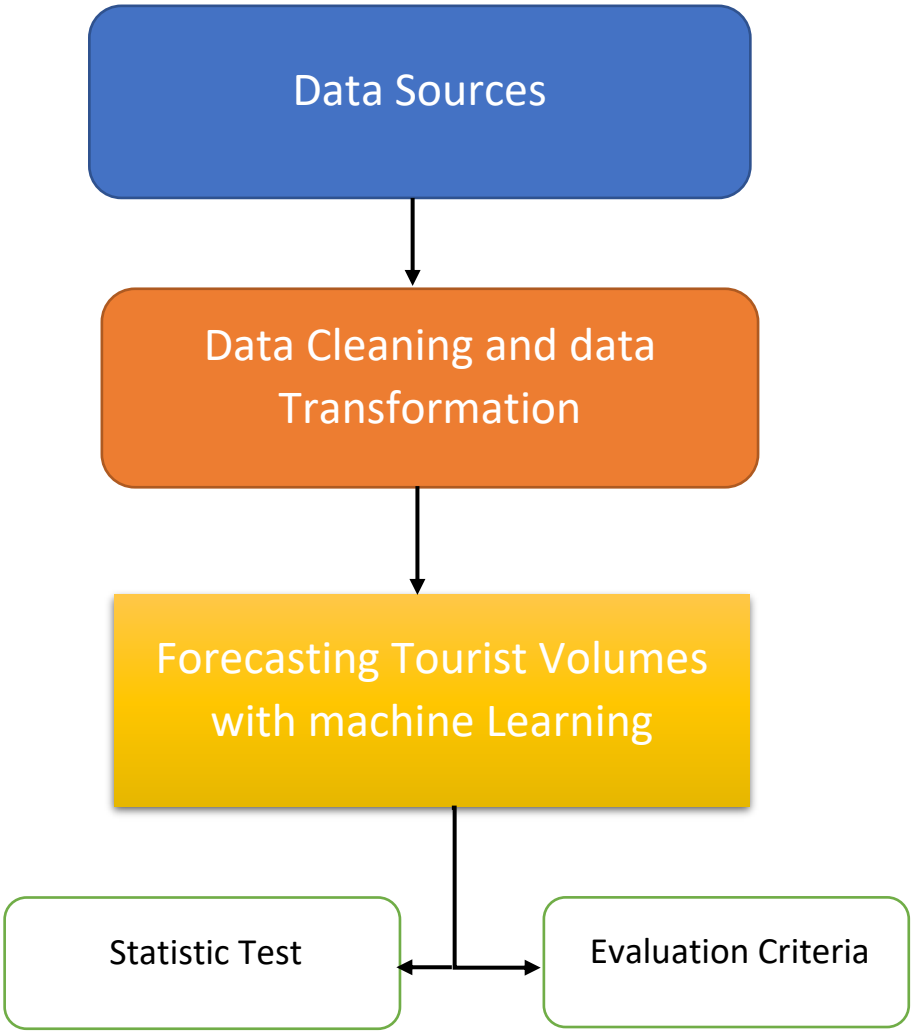


Fitting and Predicting Features





[GitHub Project Repository]



Schematic Diagram of Model

12. Scope for Enhancement

- Incorporating Collection and Updating of Real time data for features like **trend** using social media or **Internet Search Queries** for Qualitative Analysis and would help in reducing turning points (Trend change) errors.
- Performing and analysing Surveys to introduce expert opinion/Qualitative Analysis features.
- When Database size and cost becomes loosely constraining (upon development etc), then Econometric models/Causal methods **(Regression Analysis** for Example) can be implemented, which would not only help in predicting arrivals but also give insights on factors causing it.

13. Conclusion

Simplicity is often a major concern to suppliers, in which case time series forecasts are preferred. Expert-opinion techniques are useful at all levels, especially in considering and evaluating the effects of different strategies.

But this extension for small businesses is a great opportunity to improve sales and help these businesses grow. And since tourism experiences very much seasonality, forecasts should ideally be performed as an ongoing process if they are to be useful for planning purposes.

14. References

[1] Athanasopoulos, George, et al. "The tourism forecasting competition." International Journal of Forecasting 27.3 (2011): 822-844.

[2] Sun, Shaolong, Yunjie Wei, Kwok-Leung Tsui, and Shouyang Wang. "Forecasting tourist arrivals with machine learning and internet search index." Tourism Management 70 (2019): 1-10.

[3] Tourism Forecasting: a Review of Empirical Research, PAULINE J. SHELDON, TURGUT VAR Journal o Forecasting, Vol. 4, 183-195 (1985)

[4] https://vtechworks.lib.vt.edu/bitstream/handle/10919/70961/Chapter%2015%20Hospitality%20and%20Tourism.pdf

[5] https://zenodo.org/record/4133644

[6] Frechtling, Douglas. *Forecasting tourism demand*. Routledge, 2012.