Project Airline Report

By:-Yash Bunkar

1. INTRODUCTION

1. Overview

Air passenger traffic forecast is of great importance for airlines and civil aviation authorities. For airlines, accurate forecasts play an increasingly important role in revenue management. It helps to reduce the airlines' risk by objectively evaluating the demand of the air transportation business. For civil aviation authorities, air passenger traffic forecast provides a concrete basis for planning decisions in air transport infrastructure. Predicting as accurate as possible the passenger traffic for a certain airport is an aspect of major importance for both the airport management and the airline companies. The airport management must have quality forecasts regarding the future passenger flow in order to be able to properly decide regarding the future investments in the airport infrastructure, personnel policy and airport tariff policy. Moreover, airline companies use traffic forecasts to develop their strategy of operating new routes on certain destinations, to adapt their flight frequencies for the destinations operated and to establish their price policy. Thus, it is essential that the model used for predicting passenger traffic generates accurate results because these are further used for optimal allocation of financial resources in various airport investments.

2. Purpose

The main objective of this project is to build a prophet time series model that forecasts the passenger traffic for a given date. Predicting as accurate as possible the passenger traffic for a certain airport is an aspect of major importance for both the airport management and the airline companies. The theoretical quality of the forecasting models for air traffic of passengers is fundamental for obtaining the most accurate predictions. In this regard, a two-step process was used in developing the traffic forecasting model: (1) Identifying the proper regression model for traffic estimation based on the number of aircraft departures, and (2) Forecasting the number of aircraft departures for the current routesPredicting as accurate as possible the passenger traffic for a certain airport is an aspect of major importance for both the airport management and the airline companies. The theoretical quality of the forecasting models for air traffic of passengers is fundamental for obtaining the most accurate predictions. In this regard, a two-step process was used in

developing the traffic forecasting model:

(1) Identifying the proper regression model for traffic estimation based on the number of aircraft departures, and

(2) Forecasting the number of aircraft departures for the current routes

2. LITERATURE SURVEY

1. Existing Problem

The theoretical quality of the forecasting models for air traffic of passengers is fundamental for obtaining the most accurate predictions. Existing models for predicting passenger air traffic are divided into several categories:

- **Parametric models (econometric models).**

  Multiple regression models built on the relationship between passenger traffic and economic, social, demographic variables are intensely used as well as gravity models. Gravity models' premise is that the passenger is informed and acts rationally from an economic point of view. In this regard, the factors that influence the passenger's behavior to choose a particular airport are: arrival time at the airport, frequency of flights to specific destinations, and flight ticket cost (Brian Graham, 1999).

This category also includes the traffic forecasting methodology used by IATA, which estimates the market air demand based on the socio-economic variables of the market, GDP and adjustment factors: regulations, demand, airlines, competition, substitution (Air Traffic Forecasting Methodology by IATA). It has been concluded that there is generally unidirectional Granger causality from GDP to Revenue passenger Kilometers (RPK), also used as airline traffic (E. Fernandez, R.R. Pacheco, 2010). Wenbin Wei and Mark Hansen have used a logarithmic function for estimating the demand of passenger air traffic based on the following independent variables: flight frequencies, number of seats in the aircraft, ticket price, flight distance, number of spokes in the network, airport capacity, airport area population income, number of local passengers who travel from spoke S to hub H by airline A, total number of initiated passenger trips originating from spoke S (Wenbin Wei, Mark Hansen, 2006). A similar model, called the Econometric Dynamic Model (EDM) was used to predict the number of passengers according to economic variables, active population, consumer price index, number of flights, average occupancy rate of hotels, value foreign currency exchange rates values at international arrivals (Rafael Bernardo Carmona-Benítez, Maria Rosa Nieto, Danya Miranda, 2016). The mixed multinomial logit model (MMNL) has also been used with good results to analyze air travel behavior in airport

choice (Stephane Hess, John W. Polak, 2005). Fuzzy regression analyses is used for estimating passenger air traffic, but also for forecasting cargo air traffic in order to reduce the residual value resulting from the influence of ncertain and unidentified factors in parametric models (T.Y Chou, G.S Liang, T.C Han, 2011; M.S. Liao, G. S. Liang, C. Y. Chen, 2012; V.A Profillidis, 2000). Similar with the multivariate regression model is the ARIMAX model (Wai Hong Kan Tsui, Hatice Ozer Balli, Andrew Gilbey, Hamish Gow, 2014) which can mprove the forecast accuracy by considering the autocorrelation which may exist within the regression

residuals.

- **Nonparametric models, namely time series analysis**

  time series analysis eliminates the shortcomings of the parametric methods of identifying and predicting all variables that influence the passenger traffic. Univariate models of time series use only the passenger traffic past data without considering other exogenous variables. These time-sequence models generate better predictive results, according to the study by Ya-Ling Huang and Chin - Tsai Lin who used The Gray Envelope Prediction Model (GEPM) for monthly, seasonal and annual passenger traffic forecasts (Ya-Ling Huang , Chin - Tsai Lin, 2011). The Grey prediction model is
  suitable for short-term forecasts that do not have a high degree of uncertainty. A Grey - Markov model was applied by Zhang Wei, Zhu Jinfu in 2009 in order to estimate passenger traffic by integrating the Grey model (used for short periods with low uncertainty) in the Markov chain model applicable in dynamic settings.
  ARIMA autoregressive model is suitable for short term forecasts and when traffic records regular variations, cyclical seasonality, but generates significant errors when traffic variations are irregular (ACRP, 2007; M. Çuhadar, 2014; A. Danesi, L. Mantecchini, F. Paganelli, 2017). The SARIMA model is a combination of stochastic seasonal model and ARIMA model and can best describe the fluctuation of passenger flow of the airport terminal which presents a periodic fluctuation. The study carried out by Ziyu Li et al using SARIMA generated predictions of passenger traffic very close to the actual values, the error rate of the model being between 1% and 3% (Ziyu Li, Jun Bi, Zhiyin Li, 2017). Another study developed by Wai Hong Kan Tsui et al for predicting passenger traffic at Hong Kong airport using SARIMA and ARIMAX models has resulted in high accuracy, with very small prediction errors (Wai Hong Kan Tsui, Hatice Ozer Balli, Andrew Gilbey, Hamish Gow, 2014). For time series characterized by nonlinearity with irregular fluctuations and evolutions, it is recommended to use artificial neural networks (ANNs), support vector machines (SVMs), genetic programming (GP) and ensemble empirical mode decomposition (EEMD) (Y. Bao, T Xiong, Z. Hu, 2012). The LSTM (long short-term memory) network model, a neural network prediction model is proper for short-term traffic forecasting (Z. Zhao, W. Chen, X. Wu, P.C.Y. Chen, Jingmeng Li, 2017). Neural network forecasting models are also used if there is a nonlinear relationship between the models' variables (T.O. Blinova, 2007)

2. Proposed Solution

We will be building a Web application where

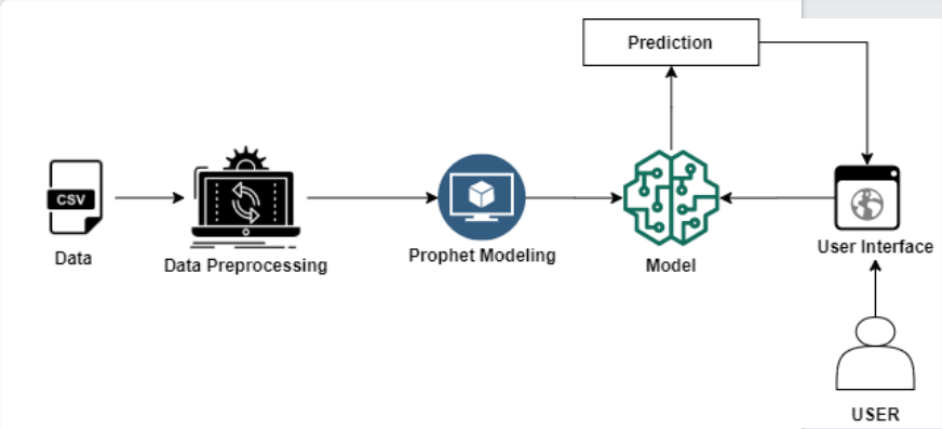
- The user selects the date from User Interface(UI)
- The passenger traffic for the selected date is analysed by the model
- The count of passengers for the selected date is displayed on UI

To accomplish this, complete all the milestones & activities listed below.

- Installation of Pre-requisites.
    - Installation of Anaconda IDE / Anaconda Navigator.
    - Installation of Python packages.
- Data Collection.
    - Create or Collect the dataset.
- Data Pre-processing.
    - Importing of Libraries.
    - Importing of Dataset & Visualisation.
- Model Building.
    - Fitting the prophet library.
    - Cross validation of the model.
    - Evaluation of the model.
    - Save the model.
- Application Development.

3. THEORITICAL ANALYSIS

1. **Block Diagram**



2. Hardware/Software Designing

Softwares are :

1. Jubiter notebook
2. Watson studio
3. Virtual studio code
4. Anaconda promt
5. Spider code

Hardwares are :

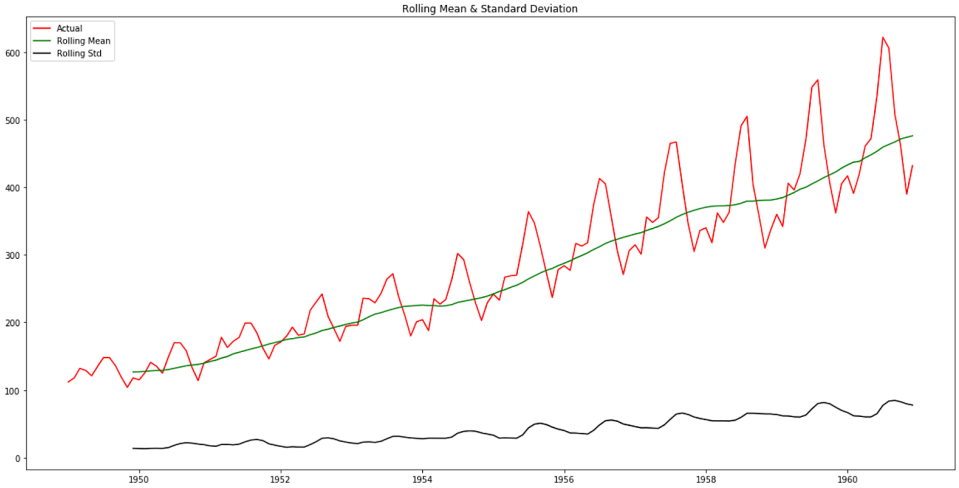
1. Windows 7 to 11
2. Ram about 8 gb
3. Intel core 5 or 7 or amd 5

4. EXPERIMENTAL INVESTIGATIONS

```
data['Month']=pd.to_datetime(data['Month'], infer_datetime_format=True)
data=data.set_index(['Month'])
print(data.head())
print(data.tail())
```

```
            Passengers
Month
1949-01-01         112
1949-02-01         118
1949-03-01         132
1949-04-01         129
1949-05-01         121
            Passengers
Month
1960-08-01         606
1960-09-01         508
1960-10-01         461
1960-11-01         390
1960-12-01         432
```

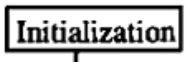
```
plt.figure(figsize=(20,10))
actual=plt.plot(data, color='red', label='Actual')
mean_6=plt.plot(rolmean, color='green', label='Rolling Mean')
std_6=plt.plot(rolstd, color='black', label='Rolling Std')
plt.legend(loc='best')
plt.title('Rolling Mean & Standard Deviation')
plt.show(block=False)
```

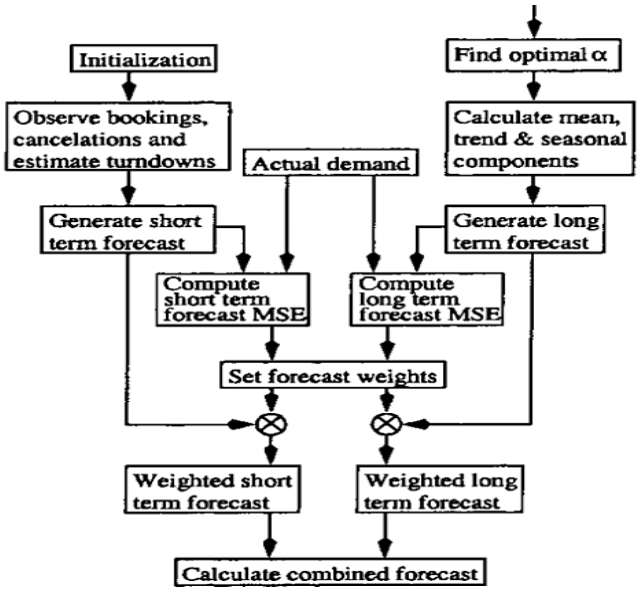



```
from statsmodels.tsa.stattools import adfuller
print('Dickey-Fuller Test: ')
dftest=adfuller(data['Passengers'], autolag='AIC')
dfoutput=pd.Series(dftest[0:4], index=['Test Statistic','p-value','Lags Used','No. of Obs'])
for key,value in dftest[4].items():
    dfoutput['Critical Value (%s)'%key] = value
print(dfoutput)
```

```
Dickey-Fuller Test:
Test Statistic            0.815369
p-value                   0.991880
Lags Used                13.000000
No. of Obs              130.000000
Critical Value (1%)      -3.481682
Critical Value (5%)      -2.884042
Critical Value (10%)     -2.578770
dtype: float64
```

5. FLOWCHART





6. RESULT

For airlines, accurate forecasts play an increasingly important role in revenue management. It helps to reduce the airlines' risk by objectively evaluating the demand of the air transportation business. For civil aviation authorities, air passenger traffic forecast provides a concrete basis for planning decisions in air transport infrastructure.

7. ADVANTAGES & DISADVANTAGES

2. Advantages

By **accurately forecasting demand for each flight or seat**, revenue management adjusts pricing to maximise unit revenue. By predicting when demand is high and relatively inelastic, lower fares can be restricted; by anticipating when demand is low but elastic, lower fares can be made more available.

3. Disadvantages

It is normally accepted that even though there are difficulties associated with making forecasts of air transport demand, estimates are necessary to:

1. Assist manufacturers in industry to anticipate levels of aircraft orders and to develop new aircraft.
2. Aid airlines in their short and long term planning for equipment, facilities and personnel.
3. Assist central governments to facilitate the orderly development of the national and international

airways system

4. Aid all level of government and airport authorities in the planning of airport infrastructure including terminal facilities, access routes, runways, taxiways, aprons, support facilities and terminal air traffic control.

8. APPLICATIONS

forecasting is **a key tool for decision making**, enabling anticipation to make short term decisions and how to respond to them as well as supporting longer term decisions with regard to future patterns in demand for air travel.

Forecasting is an important factor to any businesses because it gives the ability to make informed business decisions and establish data-driven strategies. It allows companies to have the capability to decide on the strategic allocation of resources, decide the need for corrective actions, or to adjust current strategies to reflect new situations. It is vital for companies to be always planning ahead to ensure the readiness to take on future demands as well as challenges and this is no different for companies in the aviation industry, particularly in this pandemic crisis downturn.

CONCLUSION

Passenger traffic is one of the most important factors which guides the strategic and tactical decisions of both the airport and airline company management. In this regard, accurate estimation of passenger traffic will optimize an airport's financial planning for the periods to come. The present paper describes a prediction model of passenger traffic.

9. FUTURE SCOPE

Using this, we will be able to understand:

- How Prophet can be used to make time series forecasts
- How to analyse trends and seasonal fluctuations using Prophet
- The importance of changepoints in determining model accuracy

10. BIBLIOGRAPHY

- Additive and multiplicative seasonality — can you identify them correctly?
- data.world: Air Traffic Passenger Data (Original Source: San Francisco Open Data)
- GitHub: facebook/prophet
- Airport Cooperative Research Program (ACRP), (2007). Airport Aviation Activity Forecasting.
- Antonio Danesi, Luca Mantecchini and Filippo Paganelli, (2017). Long-Term And Short-Term

APPENDIX

1. Source Code

 # STEP 1: DATA COLLECTION

 # STEP 2: DATA PREPROCESSING/ DATA WRANGLING

IMPORTING LIBRARIES

```
import numpy as np
import pandas as pd

import os, types
import pandas as pd
from botocore.client import Config
import ibm_boto3

def __iter__(self): return 0

# @hidden_cell
# The following code accesses a file in your IBM Cloud Object Storage. It includes your credentials.
# You might want to remove those credentials before you share the notebook.

if os.environ.get('RUNTIME_ENV_LOCATION_TYPE') == 'external':
    endpoint_8a2f5a52fdef4572adb9fb8e941ad5b9 = 'https://s3.us.cloud-object-storage.appdomain.cloud'
else:
```

```
    endpoint_8a2f5a52fdef4572adb9fb8e941ad5b9 = 'https://s3.private.us.cloud-object-storage.appdomain.cloud'

client_8a2f5a52fdef4572adb9fb8e941ad5b9 = ibm_boto3.client(service_name='s3',
    ibm_api_key_id='pfXou56jYL5FPwN_OE9nLk7epMJ8v0nu0RhSr6JIP0Sk',
    ibm_auth_endpoint="https://iam.cloud.ibm.com/oidc/token",
    config=Config(signature_version='oauth'),
    endpoint_url=endpoint_8a2f5a52fdef4572adb9fb8e941ad5b9)

body = client_8a2f5a52fdef4572adb9fb8e941ad5b9.get_object(Bucket='forecastcommuters-donotdelete-pr-4qmd5cequ3awqq',Key='air_passengers.csv')['Body']
# add missing __iter__ method, so pandas accepts body as file-like object
if not hasattr(body, "__iter__"): body.__iter__ = types.MethodType( __iter__, body )

dataset = pd.read_csv(body)
dataset.head()

dataset.tail()
dataset.head(10)
dataset['year'] = pd.DatetimeIndex(dataset['Month']).year
dataset['month'] = pd.DatetimeIndex(dataset['Month']).month
#dataset['day'] = pd.DatetimeIndex(dataset['Date']).day
dataset.head()
dataset.corr()
```

Dropping Date Column, Because we already splitted the Date into Year, Month and Day

```
dataset.drop('Month', axis=1, inplace=True)
```

Checking for null values

```
dataset.isnull().any()
dataset.info()
```

HANDLING MISSING VALUES

```
import matplotlib.pyplot as plt
plt.bar(dataset['month'],dataset['y'],color='green')
plt.xlabel('Month')
plt.ylabel('y')
plt.title('FORECAST COMMUTERS INFLOW FOR AIRLINE INDUSTRY')
plt.legend()
import seaborn as sns
```

```
sns.lineplot(x='year',y='y',data=dataset,color='red')
fig=plt.figure(figsize=(8,4))
plt.scatter(dataset['year'],dataset['y'],color='purple')
plt.xlabel('Month')
plt.ylabel('Price')
plt.title('PRICE OF NATURAL GAS ON THE BASIS OF MONTHS OF A YEAR')
plt.legend()
import seaborn as sns
sns.pairplot(dataset)
plt.show()
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
dataset['colum_name'] = le.fit_transform(dataset['column_name'])
```

IT IS NOT NECESSARY TO APPLY LABEL ENCODING AND ONE HOT ENCODING AS THE DATASET DOES NOT CONTAIN ANY TEXTUAL DATA

SPLIT DATASET INTO INPUTS AND OUTPUTS

Now, Split the dataset into X(independent variable) and Y(dependent variable)

```
x=dataset.iloc[:,1:3].values #inputs
y=dataset.iloc[:,0:1].values #output price only
x
y
```

SPLIT THE DATA INTO TRAIN AND TEST SETS

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,
                    test_size=0.2,random_state=0)
x_train.shape
x_test.shape
```

STEP 4: BUILDING AND TESTING THE MODEL

MULTIPLE LINEAR REGRESSION

```
#importing linear regression from scikit learn library
from sklearn.linear_model import LinearRegression
#mlr is object of LinearRegression
mlr=LinearRegression()
#trainig the model using fit method
mlr.fit(x_train,y_train)
```

y_pred=mlr.predict(x_test)

y_pred

y_test

from sklearn.metrics import r2_score

accuracy=r2_score(y_test,y_pred)

accuracy

Note: The accuracy obtained using the Multilinear Regression Algorithm is very low...Therefore we will not use this algorithm

DECISION TREE REGRESSOR

#import decision tree regressor

from sklearn.tree import DecisionTreeRegressor

dtr=DecisionTreeRegressor()

#fitting the model or training the model

dtr.fit(x_train,y_train)

PREDICTION

y_pred=dtr.predict(x_test)

y_pred

y_test

ACCURACY EVALUATION

from sklearn.metrics import r2_score

dtraccuracy=r2_score(y_test,y_pred)

dtraccuracy

RANDOM VALUE PREDICTION

dataset.head()

y_p=dtr.predict([[2005,12]])

y_p

y_p=dtr.predict([[1997,1]])

y_p

import pickle

pickle.dump(dtr,open('commuters.pkl','wb'))

pwd

 # Deployment

URLS Dallas: https://us-south.ml.cloud.ibm.com, London - https://eu-gb.ml.cloud.ibm.com, Frankfurt - https://eu-de.ml.cloud.ibm.com, Tokyo - https://jp-tok.ml.cloud.ibm.com

Import and Install dependencies

!pip install -U ibm-watson-machine-learning

```
from ibm_watson_machine_learning import APIClient
import json
import numpy as np
# Authenticate and Set space
wml_credentials = {
   "apikey":"vExgq1JGwJO_eoV2RFG2_jIatGbr8MUBLetzyQBS6kF3",
   "url":"https://us-south.ml.cloud.ibm.com"
}
wml_client = APIClient(wml_credentials)
wml_client.spaces.list()
SPACE_ID="debbba3d-0e08-45b5-a0df-f6a752e47338"
wml_client.set.default_space(SPACE_ID)
wml_client.software_specifications.list()
# Save and Deploy Model
import sklearn
sklearn.__version__

MODEL_NAME = 'forecastModel'
DEPLOYMENT_NAME = 'forecast_deploy'
FCI_MODEL = dtr
# Set Python Version
software_spec_uid = wml_client.software_specifications.get_id_by_name('default_py3.8')

# Setup model meta
model_props = {
   wml_client.repository.ModelMetaNames.NAME: MODEL_NAME,
   wml_client.repository.ModelMetaNames.TYPE: 'scikit-learn_0.23',
   wml_client.repository.ModelMetaNames.SOFTWARE_SPEC_UID: software_spec_uid
}
#Save model
model_details = wml_client.repository.store_model(
   model=FCI_MODEL,
   meta_props=model_props,
   training_data=x_train,
   training_target=y_train
)
model_details
```

```
model_uid = wml_client.repository.get_model_uid(model_details); model_uid
wml_client.connections.list_datasource_types()
# Set meta
deployment_props = {
   wml_client.deployments.ConfigurationMetaNames.NAME:DEPLOYMENT_NAME,
   wml_client.deployments.ConfigurationMetaNames.ONLINE: {}
}
```

 # Deploy

```
deployment = wml_client.deployments.create(
   artifact_uid=model_uid,
   meta_props=deployment_props
)
```