CIS 5700 Final Project Paper Winter-2019

Time Series Forecasting with Python

Sumanth Pobala

Abstract

Time series forecasting is one of trending topic in the world and many businesses use this in their daily life. This Project paper introduces the concept of time series along with different models that are used for forecasting the time series data. We will learn about how to implement different models and what role does it play in modeling time series data.

Introduction

Time series is a series of observations which are collected in time order or in equal intervals of times is known as time series. Most companies use time series data to analyze the next year's sales numbers, website traffic, traffic counts, calls received numbers, etc. For forecasting, data from a time series may be used. There are other kinds of time series data like DevOps monitoring data, mobile application event streams and scientific measurements. Time-Series data sets have 3 common things:

1. The arrival of the data is almost always recorded as a new entry.
2. The data will be in time-order
3. The data usually reaches the primary axis of time.(the times may be either regularly or irregularly).

It has equally spaced intervals like Daily, Quarterly, Monthly and Yearly.

Components of Time Series

Components of Time-Series are Long-term movement or Trend, Periodic Variations and Irregular Variations.

Trend

Trend shows a common tendency of the data to decrease or increase during a long-term movement. Trend may be upward or downward or be in a stable position. If we plot the time series values on a graph in accordance with time t. The pattern of the data clustering shows the type of trend. If the set of data cluster more or less round a straight line, then the trend is linear otherwise it is non-linear (Curvilinear). [1]

Periodic Variation

In this type of time series, the observations are repeated in a certain amount of time. There are two types of variations which are Seasonal variations and cyclic variations. If the variations occur in regular intervals of time which is less than a year like Weekly, Hourly, Monthly, Quarterly then it is a seasonal variation. If the variation is more than a year it is called Cyclic variations.

Irregular Variation

When the variations of time observations are random or irregular, it is called Irregular Variation. These are Unpredictable, Uncontrollable, Unforeseen.

Problem Statement

Unicorn Investors wants to make an investment in a new type of transportation called Jet-Rail, It uses Jet propulsion technology to run the rails and make transportation faster. The investment will make sense only, if they can get more than 1 Million monthly users with in next 18 months. To help Unicorn Ventures in making their decision, we need to forecast the traffic on Jet-Rail for the next 7 months.

Techniques used

In this project we have used few techniques and various modelling techniques to find the best model useful for forecasting the jet-rails time-series data. The techniques used are given below

1. Feature Extraction
2. Exploratory Data Analysis
3. Modelling Techniques
    - 3.1. Naïve Bayes Forecast
    - 3.2. Moving Average Method
    - 3.3. Simple Exponential Smoothing
    - 3.4. Holts trend linear model
    - 3.5. Holts winter model
    - 3.6. ARIMA model
    - 3.7. SARIMAX model

The main intention to use these models is to find the best model and use it to forecast the data. We will find the best model based on the model's Root Mean Square Error(RMSE) score.

Hypothesis

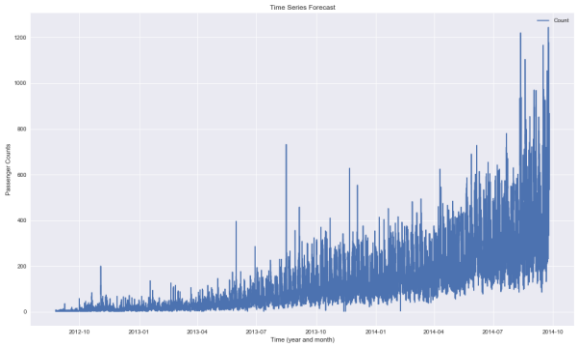
To find all the possible factors that will affect outcome, we will generate hypothesis. The following are the cases

- As the time passes, traffic will be increased as population increase year by year.
- From May to October, the traffic will be high as tourists will be visiting more in this period.
- Traffic in the peak hours will be more

- Traffic in weekdays is less than weekends

Feature Extraction

In the dataset we have object data type for date time. We will convert it into DateTime format where without conversion Feature Extraction is merely impossible. After conversion the plot for time series data will look like



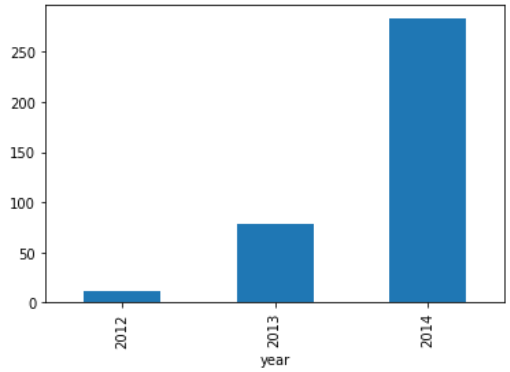
Plot 1 Feature Extraction

Even here we can observe that there is an increasing trend, the number of passengers is increasing with respect to time. We can see that at some points there is an sudden increment in the traffic where it may be occurred due to some event.

Exploratory Data Analysis

To verify our hypothesis we will perform exploratory data analysis using the actual data and we can visualize main characteristics of data using EDA.

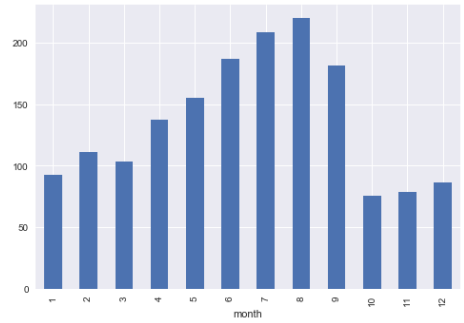
1. In the first Step we will perform EDA to look at passenger count yearly wise using groupby year.



Plot 2 Yearly samples EDA

We can see that passengers are increased as years are passing by, where our first hypothesis is verified.

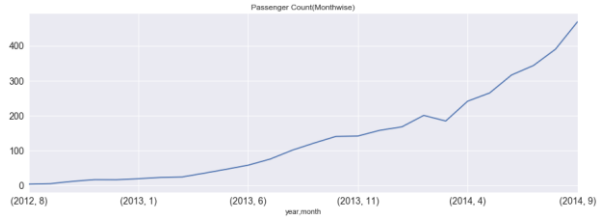
2. In the second step we will be looking at the passenger count monthly wise, where the graph looks like



Plot 3 Monthly Samples EDA

We can see that from May to October, the passenger count is high which resembles that traffic is high from May to October.

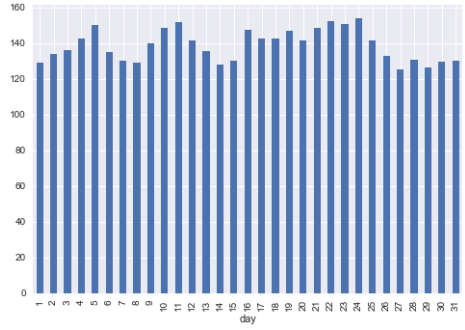
3. In third stage the graph is plotted between year, month and passenger count where the plot looks like



Plot 4 Year, Month vs Passenger count

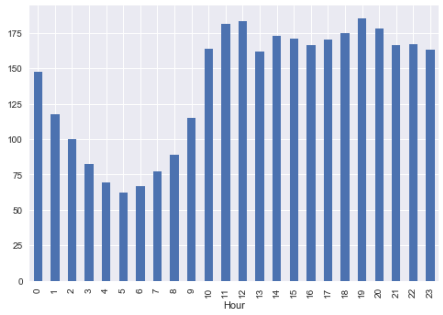
Through this plot we can see that there is an increasing trend in the passenger count and growth is approximately exponential

4. At last we will look the daily passenger count which doesn't produce much insight of data



Plot 5 Day Samples EDA

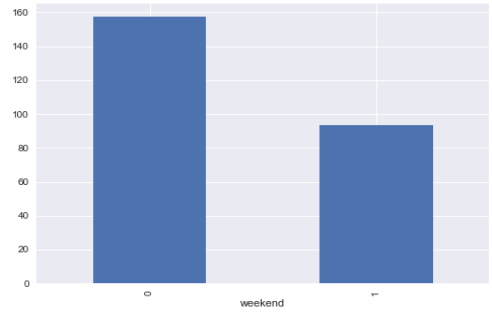
5. in hypothesis we had a statement that traffic will be high in peak hours, lets verify that by plotting hourly passenger count.



Plot 6 Monthly Samples EDA

We can see that from 6AM to 7PM there is increasing

Traffic where our hypothesis is verified. And lastly we will look the plot which visualizes the traffic in Weekdays and Weekends.

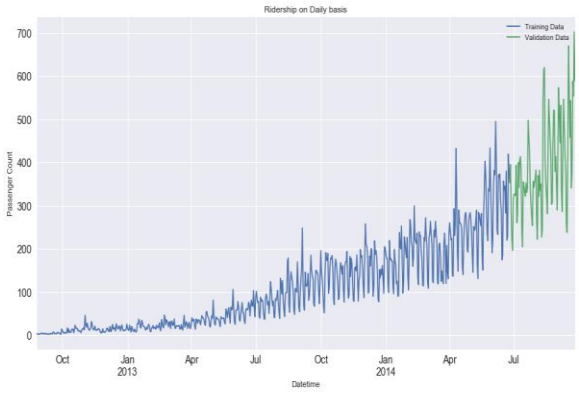


Plot 7 Hourly Samples EDA

Where 0 represents weekdays and 1 Represents Weekends and we can observe that weekdays has more traffic than weekends.

Splitting Data

After the Exploratory data analysis the data is being split into Training data and Validation data where Training data is from '2012-08-25':'2014-06-25' and Validation data is from '2014-06-25':'2014-09-25'. Now we will find the ridership based on daily basis.



Plot 8 Training Data and validation data

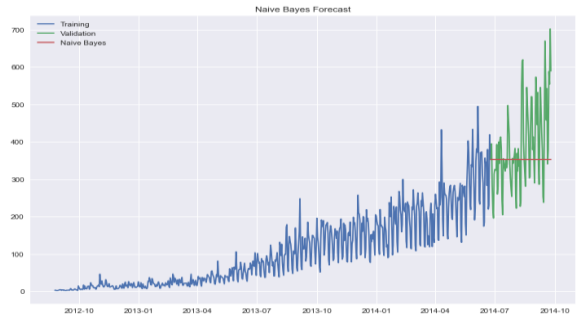
In the plot blue region shows the training data and green region shows validation data, This step is essential to build models. Last three months is selected as validation data and rest comes under training data.

Modelling Techniques

To forecast the timeseries data we will use different models

Naïve Approach

It is an estimating technique where we assume that next possible observation is equal to last observed value so that it results in a straight line which is the prediction. We store the *Training.count* in an array where the plot looks like. Where the prediction is shown in red line in the plot.



Plot 9 Naïve Approach

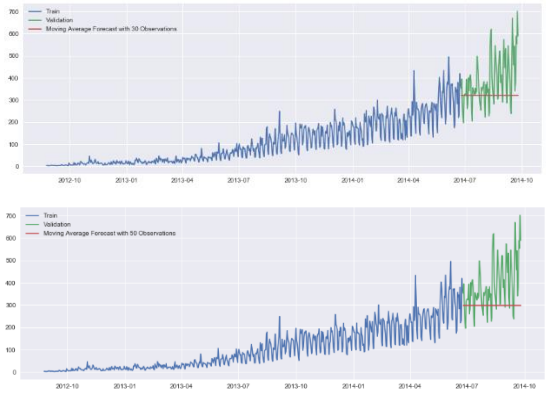
To know how accurate is our model we will calculate accuracy using Root Mean Squared Error (RMSE) score, where RMSE mean standard deviation of the residuals

$$RMSE = SQRT\sum i=1N1N(p-a)2 \text{ ------(1)}$$

In python we will import mean squared error from a package called sklearn.metrics and calculate the rmse where *RMSE score for Naïve approach is 116.07386*. Naïve approach is not suitable for the dataset which has high variability and this technique resulted in a high error score. We will adopt different techniques to reduce the rmse score.

Moving Average Method

After Naïve approach we are going to apply Moving Average Method where average of the passenger counts for few time periods is taken into consideration. This method is used to get an overall idea of the trends in the datasets. This method is extremely useful on forecasting the long-term periods. In the project we will apply Moving Average for 20, 30, 50 observations. The observation here is as the number of observations are increasing, the average is getting decreased. A snapshot of the plot with 30 and 50 observations is provided below.



Plot 10 Moving Average with 30 and 50 Observations

You can observe that the red line which is prediction with moving average is decreased. This method produced an RMSE of 142.78 which is higher than Naïve Approach

Drawbacks of Moving Average method [2]

1. With this method forecasting the future values is not possible where it is important.

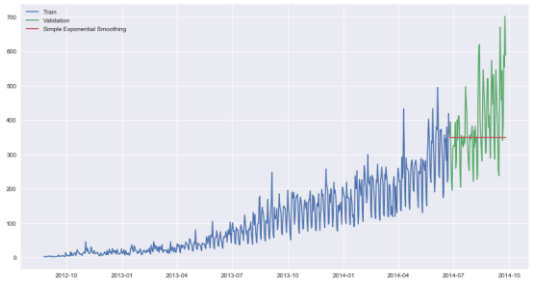
2. It does provide the trend values for all terms

3. The assumption in this method is that trend is linear, but it is not always the case.

4. The main problem is to determine the extent of the moving average which completely eliminates the oscillatory fluctuations.

Simple Exponential Smoothing

In this method assigning weights concept comes into place where larger weights are assigned to recently observations and weights exponentially decrease as we go into past values. The smaller weights are associated to old observations. This method is similar to the Naïve approach as we are caring only about the recent observations. In this project we used a smoothing_level of 0.4 and optimization as True. The plot is given below.



Plot 11 Simple Exponential Smoothing

The RMSE for this model is 117.723 which is better when compared to moving average and it is improved now.

Advantages of Exponential Smoothing

1. It produces a bit more accurate forecasts using the trend projection techniques.

1. Like Naïve approach, simple exponential smoothing also gives more significance to the recent observation.

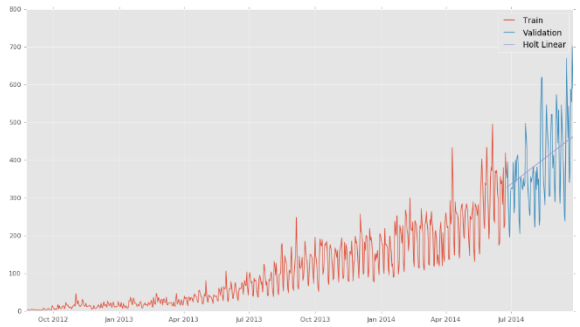
Drawbacks of this model

1. It is not useful when cyclic or seasonal variations are present as it is used to forecast short term movements.

2. As it uses a concept called smoothing it produces a forecast which lags behind the actual trend [3]

Holts Trend Linear Model

This method is applied on daily time series and after predicting the daily time series we will predict the hourly samples. We will fit the holt trend model on training data and validate it using the validation set and late we will make predictions using the test set.



Plot 12 Holts Trend Linear Model

We can see in the plot that there is a slant prediction line directed upwards which is Holts Linear trend line prediction and it produced an RMSE of 96.49 which is much better than any other models as this model considers the trend in data and forecasts the data well.

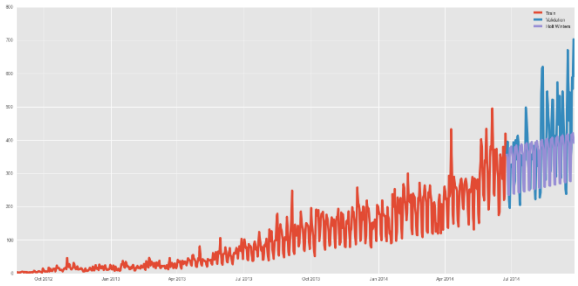
As this model produced a good accuracy now we will predict the test set and export the prediction into an excel file. The output would look like below where the attribute count is actual prediction observations.

	ID_x	prediction	ID_y	Hour	ratio	Count	ID
0	18299.5	328.12898	18288	0	0.044287	348.761110	18288
1	18299.5	328.12898	18289	1	0.035343	278.329612	18289
2	18299.5	328.12898	18290	2	0.029911	235.552579	18290
3	18299.5	328.12898	18291	3	0.024714	194.628681	18291
4	18299.5	328.12898	18292	4	0.020802	163.819549	18292

Output 1 Holts Trend Model Output

Holts Winter model

In time series the common issue is datasets suffer from seasonality due to the similar patterns or fixed intervals. Holts Trend method doesn't consider Seasonality into account. We need a model which considers Seasonality into the consideration and forecasts the data. So we took Holts winter model which forecasts with seasonality data. After applying the data it produced an RMSE of 111.40 and the it has increased again. We will follow the same method like holts trend method to predict the test data and export it to new excel file. The output of Holts winter method look like the below plot.



Plot 13 Holts winter model

Where Red region indicates Training data, blue indicates validation data and purple is the Holts Winter model prediction.

It produced a increased RMSE because it didn't work well with the seasonality data and it worked well with trend data.

ARIMA MODEL

All the above models either worked with trend data or seasonality data, we need a model which works with both trend and seasonality data, one such model is ARIMA model.

ARIMA mean Auto Regressive Integration Moving Average which is a combination of Auto Regression and Moving Average which is AR + I + MA where I stands for Integration. It has three parameters p (Order of Auto regressive model), d(Degree of differencing) and q(Order of moving average model).

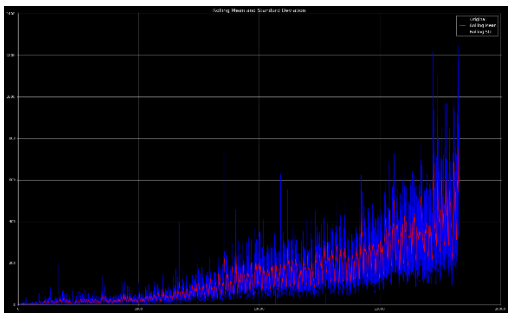
Forecasting for a stationary time series with ARIMA model is nothing but a linear equation. If mean and variance of the time series shouldn't be a function of the time then it is stationary time series. It should be constant. We check for the stationarity in time series, if it is not, we will convert into stationary time series. We do it because we will make the variables independent, When variables are independent we will get more information. We can remove the stationary by removing the trend and seasonality from the data.

Parameter Tuning for ARIMA

We will use Dickey Fuller Test to check the stationarity of the series which determines how strong a timeseries is dependent on the trend.

We will have null hypothesis as time series is not stationary and alternative hypothesis as it is stationary.

The result of dickey-fuller test is



Plot 14 Parameter Tuning

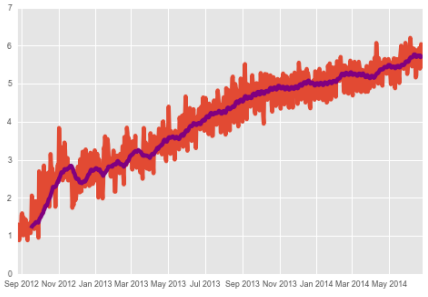
```
Results of Dickey Fuller test:
Test Statistics                  -4.456561
p-value                           0.000235
# Lag Used                      45.000000
Number of Observations Used    18242.000000
Critical Value (1%)              -3.430709
Critical Value (5%)              -2.861698
Critical Value (10%)             -2.566854
dtype: float64
```

Figure 1 Dickey-Fuller test 1

We can see that Test-Statistics is less than critical values which says that time series is stationary.

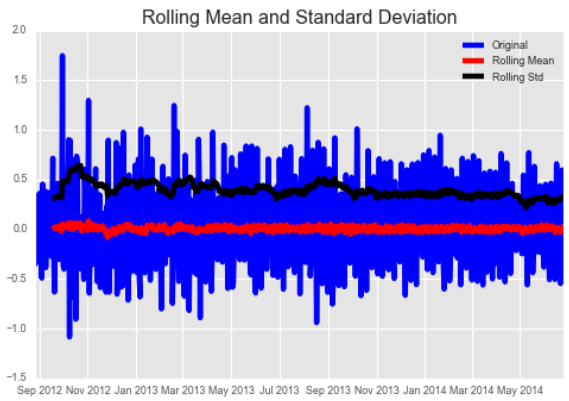
Removing Trend

We need to remove trend because we can observe an increasing trend previously We can remove the trend using the log transformation where higher values suffer more than smaller values and by rolling average where window size will be 24 as there are 24 hours in a day.



Plot 15 Removing Trend

We can see an increasing trend in the above picture. To make the series stationary we will remove this trend. Now we will stabilize the mean of the series with differencing which will result in the following plot.



Plot 16 Removing Trend

Here Black line is Rolling Standard Deviation and Red line is Rolling Mean. The results are given below

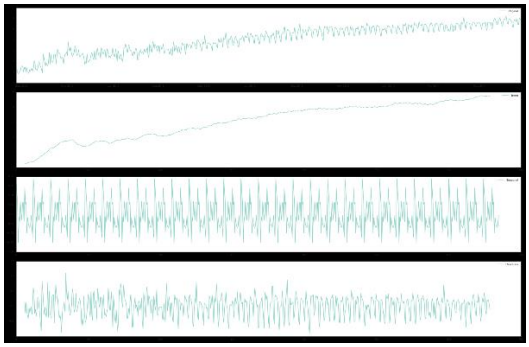
```
Results of Dickey Fuller test:
Test Statistics                -8.253359e+00
p-value                         5.317209e-13
# Lag Used                      1.900000e+01
Number of Observations Used     6.490000e+02
Critical Value (1%)            -3.440466e+00
Critical Value (5%)            -2.866004e+00
Critical Value (10%)           -2.569147e+00
dtype: float64
```

Figure 2 Dickey-Fuller test Removing trend

Where Test Statistics is much smaller than Critical Values so it is verified that trend is removed.

Removing Seasonal Data

We need to remove the periodic fluctuations because a time series is influenced by seasonal factors. We will use the seasonal decompose the time series into trend, seasonality and residuals



Plot 17 Removing Seasonality

In this we will check the stationarity of residuals.

```
Results of Dickey Fuller test:
Test Statistics                -7.830840e+00
p-value                         6.299419e-12
# Lag Used                      2.000000e+01
Number of Observations Used     6.250000e+02
Critical Value (1%)            -3.440856e+00
Critical Value (5%)            -2.866175e+00
Critical Value (10%)           -2.569239e+00
dtype: float64
```

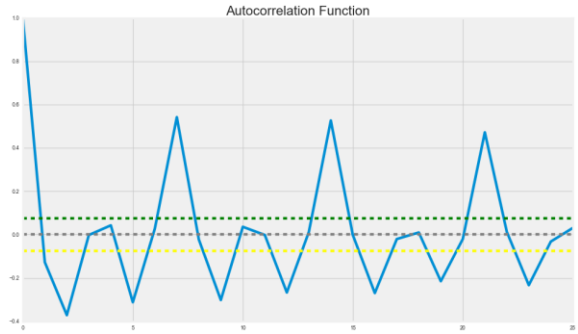
Figure 3 Dickey-Fuller test Checking Residuals stationarity

As the Test statistics is less than critical values we can say that seasonality is removed.

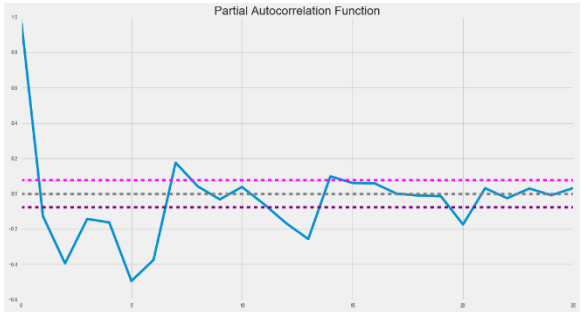
Forecasting using ARIMA

Firstly we need to optimize the p,d,q parameters which we can do it with AutoCorrelation Function(ACF) and Partial AutoCorrelation Function(PACF)

ACF is generally correlation between a lagged version of time series and time series. PACF does the same but after removing the variations.



Plot 18 ACF plot

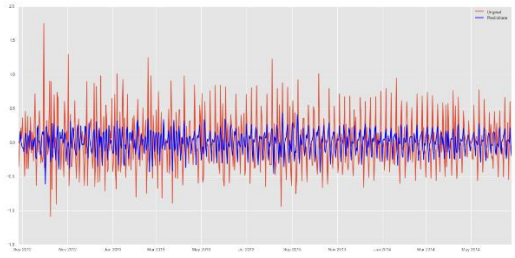


Plot 19 PACF plot

We can take now p as 1 and q as 1 with the reference of graphs.

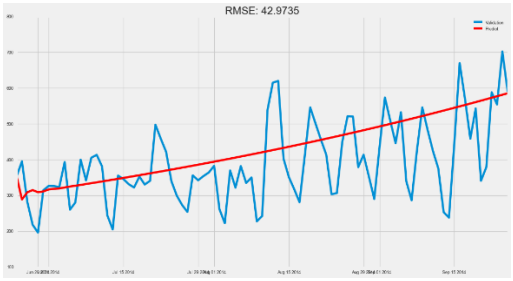
AR Model

It is used to show that output variables are linearly dependent on its own previous values. P,d,q values are 2,1,0.



Plot 20 AR Model

Validation curve for the AR Model is

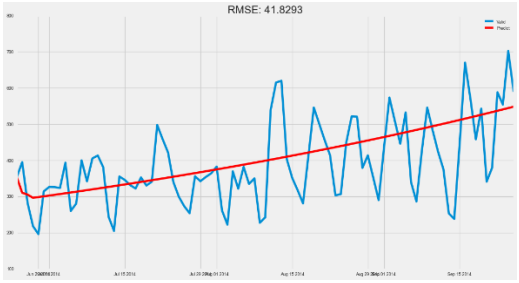


Plot 21 Validation Curve for AR Model

The RMSE score is 42.9735 .

MA Model

This model shows that output variables are linearly dependent on current values. The validation curve for the MA Model is

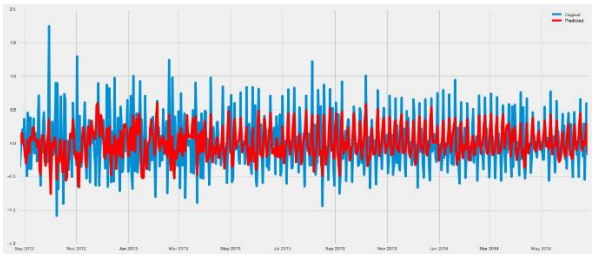


Plot 22 Validation Curve MA model

It produced an RMSE of 41.8293. The P, d, q parameters are 2, 1, 0.

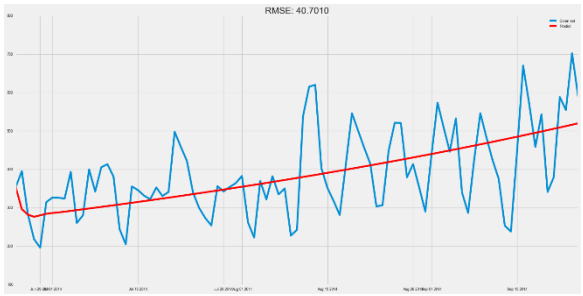
Combined Model

Now we got both models and their accuracies, we will build a combined model which makes the model stable. The parameter values p, d, q are 2, 1, 2.



Plot 23 ARIMA model

The validation curve here looks like

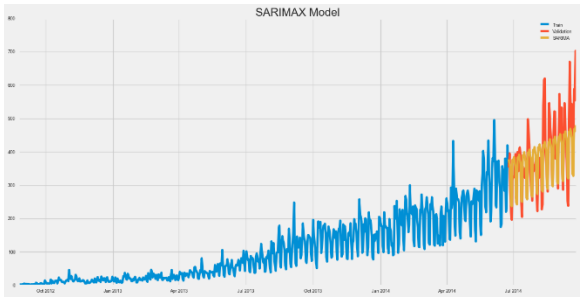


Plot 24 Validation curve for ARIMA model

Finally ARIMA model produced an ***RMSE of 40.7115*** which is better than all the above models. ARIMA model produced a great result because we removed the seasonality and trend from the time-series data which made the model is stable.

SARIMAX Model

SARIMAX model means Seasonal ARIMA exogenous Variables. It takes seasonality into accountability. The order here used is p =2, d=1, q=4 and Seasonal Order 0, 1, 1, 7, which represents the seasonal components for the difference between AR and MA models and periodicity. The resulted plot is given below where yellow represents SARIMAX prediction.



Plot 25 SARIMAX model

This Produced RMSE is 70.65 which is more than ARIMA model.

Final Output

Now we will predict on the test file and we will predict the passenger data and export it to the excel file where the final prediction looks like

	month	year	Hour_x	Datetime	Hour_y	Count	ID
5107	4.0	2015.0	11.5	2015-04-26 19:00:00	19	746.282776	23395
5108	4.0	2015.0	11.5	2015-04-26 20:00:00	20	717.151974	23396
5109	4.0	2015.0	11.5	2015-04-26 21:00:00	21	669.385484	23397
5110	4.0	2015.0	11.5	2015-04-26 22:00:00	22	671.813931	23398
5111	4.0	2015.0	11.5	2015-04-26 23:00:00	23	655.934002	23399

Output 2 Final Output

RMSE Values of all models

MODEL NAME	RMSE VALUE
NAÏVE APPROACH	142.78
MOVING AVERAGE	117.27
HOLTS LINEAR TREND	96.49
HOLTS WINTER	111.47
ARIMA	40.070
SARIMAX	70.65

Table 1 RMSE Values of all models

Applications

The applications of Time series forecasting are

- Economic Forecasting
- Stock Analysis
- Census Analysis
- Process and Quality Control
- Sales Forecasting
- Weather Forecasting

Conclusion

Through this project I have learned forecasting the time series data and found out that we have many models but only one or two models produce the efficient output where in this project ARIMA produced a better prediction because it produced a least RMSE and the model is developed after removing the seasonality and trend from the time series data. This project made me to learn many models which are used to forecast the data and how to use seasonal decompose to remove seasonality and as per the final output there are 2.7 million users in prediction and Unicorn Investors can invest in Jet-Rails as passenger count exceeded 1 million.

References

1.Trend, Components of time series, https://www.toppr.com/guides/business-mathematics-and-statistics/time-series-analysis/components-of-time-series/

2.Drawbacks of Moving Average method, https://www.toppr.com/guides/business-mathematics-and-statistics/time-series-analysis/moving-average-method/

3.Simple Exponential Smoothing, advantages and disadvantage, https://connectusfund.org/5-advantages-and-disadvantages-of-exponential-smoothing

4. Dataset for the project- https://datahack.analyticsvidhya.com/contest/practice-problem-time-series-2/

5. Theory Concepts for Time-Series-Forecasting- https://machinelearningmastery.com/time-series-forecasting/