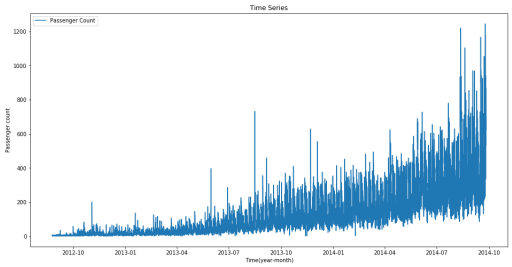
Project Report
Time Series Forecasting of Traffic on New Transportation Service

Krinza Momin
K16-3788
CS481 Data Science
BS Computer Science
FAST-NUCES, Karachi
k163788@nu.edu.pk

1 Introduction

This project aims to solve a time series problem. Time Series is generally data which is collected over time and is dependent on it and that data can be used for forecasting. The most important components of time series are Trend (general direction in which something is developing or changing) and Seasonality (pattern repeating at regular time interval) which will be analyzed in this project.



2 The Problem Statement

Some investors are interested in investing to a new transportation service to be launched soon. The investment would only make sense if there's some sort of guarantee that more than 1 million monthly users will be using this transport service within next 18 months.
In order to make decision for investors easy this project solves and forecasts the traffic for next 7 months.

2.1 The Dataset

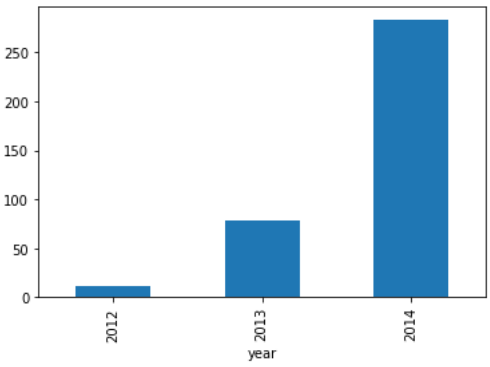
The traning and testing datasets are collected from Analytics Vidhya.
The train dataset "Train" contains three columns named ID, Datetime and Count. Whereas, Count being the label. The testing dataset "Test" contains only ID and Datetime, Count to be predicted.

2.2 Hypothesis

Some hypothesis were made before diving into the code part of this project and below are the visualizations of the hypotheses which was thought can affect the passenger count.

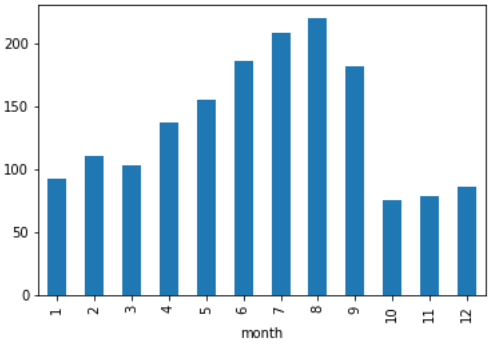
2.2.1 Hypothesis 1: Yearly passenger count will increase.

As population count increases every year, it is expected that passengers will increase too.

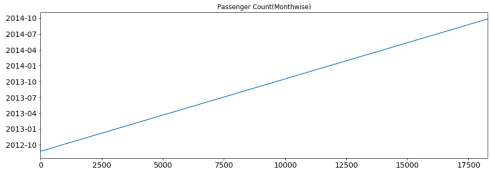


2.2.2 Hypothesis 2 : Increase in traffic from June to August.

Holiday season

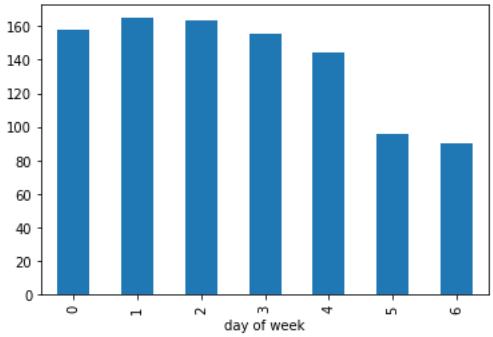


2.2.3 Hypothesis 3 : Monthwise passennger count will increase.



2.2.4 Hypothesis 4 : On weekdays passenger count will be high.

As people go to work, school and offices. At weekends less people travel.



3 The Modeling Techniques

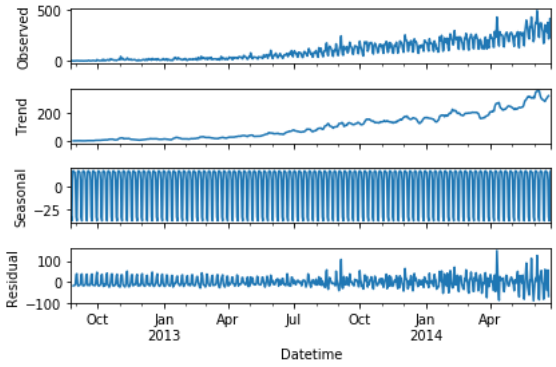
As the target variable is numerical it can be predicted using regression techniques, but a time series problem is different from a regression problem because time series is time dependent and along with an increasing or decreasing trend, most Time Series have some form of seasonality trends.

3.1 Moving Averages

In this technique the average of the passenger counts for last few time(for past 10, 20 and 50 observations) periods were taken only.
The predictions came out to be week, with the RMS value of 144.19.

3.2 Holt Linear Trend Model

The model allows forecasting of data with a trend.This method takes into account the trend of the dataset. The forecast function in this method is a function of level and trend.
Time series is decomposed in following parts as visualized:



The RMS value of Holt's linear model has decreased (Sub.csv) and therefore now we will be predicting the passenger count for the test dataset using various models. (Holt linear.csv)

3.3 Holt winterâĂŹs model on daily time series

This method takes into account both trend and seasonality to forecast future count.
The predictions came out to be really good, with the RMS value of 82.37 only.

3.4 ARIMA Model and SARIMAX

We make the series stationary to make the variables independent. Variables can be dependent in various ways, but can only be independent in one way. So, we will get more information when they are independent. Hence the time series must be stationary. SARIMAX model takes into account the seasonality of the time series. So we will build a SARIMAX model on the time series.
SARIMAX gave the least RMS value out of all and that is 70.01 only.
Check SARIMAX.csv

4 Personal Learning from this project

1. Understanding Data
2.Hypothesis Generation
3.Exploratory Analysis
4.Forecasting using Multiple Modeling Techniques
5.Visualization
6.Got started with online competitions.