Forecasting
Sports Popularity
using
Time Series Analysis

-Report by Avantika Bhattacharya

Problem Statement

Introduction

Time series analysis is an approach to forecasting commonly used in business to produce and improve point forecasts where regression falls short. Time series forecasting is increasingly in demand due to its ability to predict events based solely on previously observed data of the given event. Studies have also been done showing that early patterns found in web popularity reflect long term interest in a topic. In other business studies, search engine popularity has been shown to reflect general popularity and interest in a specific product. Our models apply this interest assumption, using major sports leagues in India and United States as our product. Forecasting has been a growing trend in the world of sports, where it has been used in an attempt to predict outcomes of games. Our analysis focuses on a separate and more general area within sports, the popularity of entire leagues. With such large market values, even small changes in future popularity could have large business implications on marketing, social media promotion, and team value. In order to model sport popularity, we pulled data from Google Trends. Google Trends is an analytical tool that allows users to compare the popularity of search terms over time. It can be used to gain insights into popularity that may not otherwise be noticed, as shown in the 2016 US presidential elections. The forecasts could make a difference for the major sports leagues' business interests due the huge amount of money involved with the leagues. Small percentage differences in popularity could mean thousands more people looking at ads, thousands more people buying merchandise, and thousands more in profits. Businesses interested in advertising or marketing to or investing with either league may find these forecasts useful for deciding which sports league provides the greater short-term or long-term value.

Objective

The objective of this project is to compare and contrast major sports league popularity using univariate time series forecasting models in order to efficiently predict the trend popularity for and between the two leagues in the Unites States and India. We wanted to make a confident prediction about which league is growing faster. We believe sport's popularity is tailor made for time series forecasting. Sports have very distinct seasons, which allow us to build a seasonality component and trend into our models.

- Source the data set from Google Trends filtering it to specific locations and time periods catering to each sports league.
- Compare the trends of rise and fall in popularity of the sports league especially in their peak seasons.
- Forecasting performance and predicting future popularity.
- Model comparisons for ARIMA and SARIMA models.
- Final concluding remarks that companies looking to invest or advertise in the respective leagues can use to make smart business decisions.

Data

Our data was sourced from the Google Trends website. This data shows how the popularity of a term has changed over time in Google searches. We looked at the specific search terms like "IPL" and "NFL". To see the scores relative to each other, we used the compare feature on the website. The data was available from November, 2007 onwards till October, 2022 at the monthly level, giving us multiple observations. We filtered the data down to searches from only the United States and India. The trends are scored using a relative index of 0-100, with 100 being the point at which the most popular term being compared peaked in popularity. A value of 50 is 50% as popular as the peak. The dataset was divided for model building, model testing and model validation purposes. Descriptive statistics and other results have been specified as well.

Procedure

- First, we sourced the dataset from Google Trends and Yahoo Finance. Then, we proceeded with cleaning the dataset and performing EDA. We observed our time series data by plotting it.

- Then we tested the dataset for stationarity using the Augmented Dickey-Fuller Test. Next, we transformed our non-stationary dataset to stationary via differencing. Then, we plotted the Autocorrelation Function and Partial Autocorrelation Function graphs.

- ACF and PACF plots are used to estimate the order of our model. ACF plot gives the order for the MA model called order q. PACF plot gives the order for the AR model called order p.

- For our IPL dataset, as we can observe from the ACF and PACF plots, the order p=3 and the order q=3. This implies that we can observe spikes in the plot at three lag values.

- Similarly, for our NFL dataset, we can observe from its ACF and PACF plots, that the order p=2 and the order q=1. This implies that the PACF plot spikes at two lag values and the ACF plot spikes at one lag value.

- We use these orders to fit our ARIMA and SARIMA models on both the datasets. ARIMA(p,d,q) where p is the order we get from the PACF plot, q is the order we get from the ACF plot and d represents the order of differencing.

- Since our initial dataset was non-stationary in nature, we have used d=1 for our models. After trying multiple combinations of orders p and q, we have chosen the below orders as their models have the lowest AIC values.

- AIC or Akaike's Information Criteria is a statistical criteria used to compare the quality of different statistical models. SARIMA (p,d,q,s) model is similar to the ARIMA (p,d,q) model with the notable addition of the seasonality (s) parameter.

- Since our dataset is a monthly time series with seasonality present in it we have set s=12. We have used the statsmodels library for ARIMA and SARIMA modelling.

- After fitting our models and summarizing them, we move towards forecasting the data. We plotted the forecasted data along with the original data to see the accuracy of the model in predicting the future trend.

- Lastly, we have performed time series analysis on the PepsiCo stock returns dataset as they are the long standing sponsors of the leagues to check for fluctuations dependent on the seasonality of the sports leagues' playing seasons.

References and Inspiration

https://www.researchgate.net/publication/319384255_Forecasting_Sports_Popularity_Application_of_Time_Series_Analysis

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwic5pK3_az6AhVi6HMBHZqpAmgQFnoECBUQAQ&url=https%3A%2F%2Fjournals.sagepub.com%2Fdoi%2F10.1177%2F1012690206063508&usg=AOvVaw2eStxjrN1rfzUeldTrMW6g&cshid=1664007120659673

https://trends.google.com/trends/explore?date=2007-09-24%202022-09-24&geo=US&q=nfl

https://trends.google.com/trends/explore?date=2007-09-24%202022-09-24&geo=IN&q=ipl

https://trends.google.com/trends/explore?date=2007-09-24%202022-09-24,2007-09-24%202022-09-24&geo=IN,US&q=ipl,nfl