**Title:- Time Series Sales Forecasting**

**Authors:-** James J. Pao, Danielle S. Sullivan

**Link To Research Paper**:-

[**http://cs229.stanford.edu/proj2017/final-reports/5244336.pdf**](http://cs229.stanford.edu/proj2017/final-reports/5244336.pdf)

**Abstract**:-

This paper talks about the dire need for accurately predicting forecast data in today’s day and age. The prediction can be related to sports performance, stock markets or even natural phenomenon. However the topic of discussion in the paper is related to forecasting retail sales. The prediction models implemented here are Regression Trees, Seasonal Trend Decomposition using Loess and Auto-Regressive Integrated Moving-Average(STL-ARIMA) and Neural Networks.

**Introduction:**-

As retail markets all around the world are going global and competition is increasing, optimization of a retail store company’s operations in terms of inventory-stocking, procurement as well as distribution of goods is very important. Forthcoming demand must be equally met with appropriate supply, i.e. potential losses occurring either due to low inventory produce or overstocking the inventory needs to be avoided.

The dataset used was that provided by Walmart, an American giant in the retail space, for Kaggle Data Science Competition. It **relates very much** to the **Kaggle dataset** that we will be using for our project. For the same, patterns from historical data are carefully monitored and inferences have been drawn.

The methods implemented for finding the solution are Regression Based Decision Trees, STL+ARIMA Model, Neural Networks, k-Nearest Neighbours(kNN), Naïve Bayes Algorithm. However, it was later found out that kNN and Naïve Bayes do not return significant insights and hence are ignored in further discussions.

**Implementation:-**

The regression based decision tree is constructed making use of the features provided in the dataset. This can be implemented pretty easily as the data is already provided the way we need it to be. This is considered as the baseline method for the project.

Apart from this, a widely used method for analysing Time Series Data is the STL+ARIMA model. The model segregates the trend(long term non-seasonal trends in the data), seasonality(periodic patterns in the data)and extra(remnant data not already captured in trend and seasonality data) components from the data. Following the decomposition of the time series data, the seasonality and trend data is extended further to test data and ARIMA also manages to include the remnant data into this.

Time Lagged Feed Forward Neural Network(FFNN) are used as a final implementation method. This a highly flexible universal approximator model. They are able to perform end-end learning with just the training data without having the need for any user-feedback.

**Assumptions:-**

The ARIMA Models have been found out to be accurately forecasting data when it takes it to be non-linear and provided there are no sudden spikes or dips in behaviour. When considered as linear data, these tend to predict poorly when the above conditions, i.e. linear data assumption and sudden changes in values, are observed.

**Main Claims:-**

The STL+ARIMA model performed better compared to the regression based decision tree. This is based on the Mean Absolute Error calculated for both wherein the STL+ARIMA got a score of 2800 and Regression based Decision Tree got a score of 4400.

This is largely down to the fact that the STL+ARIMA model does not account for the large spikes as mentioned in the assumptions.

The best FFNN model, based on varying number of hidden layers, also got a score of around 2400 and the fact that having too many hidden layers made the model overfit the data if no regularization is done during the training phase, was also discovered.

**Main Takeaway:-**

STL + ARIMA was confirmed to be a very effective model in modelling time series data with trend and seasonal components. Time-lagged FFNNs using both immediate time-lagged data as well as mean weakly sales values as inputs were also implemented

that showed very good accuracy for the best implementations(based on the number of hidden layers). This showed that neural networks can effectively model time series data and can do so without any data pre-processing such as deseasonalizing.