**YouTube Reviews Text Mining**

*Pooja Shah*

**Table of Contents**

1. **Executive Summary**
2. **Data Overview**
   1. Explaining Amazon data set
   2. Column summary
   3. Data cleansing
3. **Data Exploration**
   1. Exploring dependent variable
   2. Exploring independent variables
      1. Checking for time series
      2. Checking for cross-correlations

3.3 Checking for stationarity

**4. Modeling and forecasting**

4.1 Finalizing key independent variables

4.2 Determining right set of models

4.3 Modeling and forecasting

4.3.1 ARIMA

4.3.2 ARIMAX

**5. Model comparison**

5.1 Accuracy & fit statistics with and without hold out sample

5.2 Final Forecasting Parameters & Equation

**6. Conclusion**

**1. Executive Summary:**

The dataset contains historical sales data for 45 Amazon stores located in different geographies. Every store contains 98 departments and each of the departments has got a certain weekly sale. The data contains 3 years of data for the time horizon: 2019-02-05 to 2021-11-01 split at weekly level for 45 Amazon stores. We will be evaluating time-series forecasting to determine sales for one of these stores.

By understanding the sales trends, the store can work on optimizing inventories for lean periods and jack up the inventories where the sales would shoot up. Also, stores can run promotional events to gain an uplift in sales in case the projections show feeble sales in the coming weeks.

We have run model of ARIMA and ARIMAX with mean weekly sales as dependent variable, and temperature, fuel price, CPI, unemployment, and holiday as independent variables. We have run ARIMA(0,2), ARIMAX(2,0), ARIMAX(1,0), ARIMAX(2,2), ARIMAX(0,2), ARIMAX(0,0), ARIMAX(1,1), ARIMAX(0,1).

We found out that ARIMAX(2,0) is best model because looking at the accuracy and statistics fit, ARIMAX(2,0) has the lowest MAPE - 2.2% than any other models. In addition, it has no significant autocorrelation exists and has slightly white noise, whereas other models appear to either have autocorrelation or white noise.

In summary, ARIMAX(2,0) is optimal models for business and shows the declining of the sales trend the forward time horizon. Given that the macroeconomic indicators shows a positive trend, the forecast shows a decline in sales, hence, the recommendation for the business would be investigating the factors leading to drop; department level sales trends can be understood to find critical focus areas; and other indicators like quality, delivery of service, price points can be evaluated to understand if there are other factors influencing the sales.