ISyE 6501 Introduction to Analytics Modeling

Course Project –fall ’2021

Project Summary

Business Objective:

Demand forecasting for ABC Corp, a Logistics Management Organization

ABC Corp is looking into demand forecasting to give businesses valuable information about their potential in their current & upcoming market, so that managers can make informed decisions about pricing, business growth strategies, and market potential.

Without demand forecasting, ABC Corp, has been risking poor decisions about their products and target markets. The project focus on addressing the Concerns of negative effects on inventory holding costs, customer satisfaction, supply chain management, and profitability.

High Level Technology Objective:

Build a 12-month demand forecast at the product level

Clustering: As we are faced with the challenge of over 10k line of products for ABC Corp, best approach to our demand forecasting model is to cluster the products to groups based on their selling pattern.

Time series analysis:

ABC, being a well-established businesses who have several years’ worth of data to work from and relatively stable trend patterns available, time series analysis approach can be effectively used.

Regression:

Successful demand forecasting isn’t a one-and-done task. It involves active ongoing demand shaping by optimizing the customer experience, product offering, and inventory management. Here is where, we will be using variable selection using Elastic Net, followed by Regression to conduct ongoing analysis to identify the key predictors impacting the demands.

Approach

Step1: Data set, Collection process

8 years of Historical data is available for each product or product line.

Data set includes below attributes related to the product

Product predictors:

Product SKU | Line of Business | Unit of Business | Acquired Date | Sold Date | Product Cost | Profit Margin | Turnover rate | Inventory stock units |

External predictors:

Geographic Demographics | Weather | Major Political/Natural stress Indicators

Step2: Outlier detection and Imputation

Outlier detection: As we have time series data available, using visual outlier detection via Control chart to set thresholds at two standard deviation from median for each year to identify the Outlier.

This is followed by investigation of the root cause of the occurrence of this outlier to choose the appropriate handling technique for Outlier (Inclusion/Exclusion of Outlier). This process can be extended by looking for contextual or collective outliers considering the seasonality and cyclic patterns of the data

Imputation: As the demand forecast will most likely by impacted by recent year trends and Business' appetite, it is safe to restrict the data imputation to the most recent years. For sake of simplicity, considering 50% of data (4 years). Study the missing variables and use appropriate Imputation technique (if required) to correct the data.

Selection: Though variability is less expected on this data set, I will stay away from perturbation.

Though the preference is towards regression model to predict the Imputed value, considering this use case, we will subsequently be using Regression model to study the impact of variables to further tune and rebuilt Demand forecasting model. This will cause over fit of data if Imputation by Regression is used. Keeping these factors in mind, if no bias is detected with the missing values, I will go with removal of the missing entries to avoid error introduction

Step3: Preparation

1) Scale the data
2) Establish 70%-30% for validation and test data
3) As we will conducting factor based analysis through time series models, De-trending will be a pre–requisite to be done on both predictors and responses.
4) With focus on simple and yet effective that can be explained efficiently to business, Variable selection will be conducted using Elastic Net approach.

Step4: K Means Clustering

With the assumption that data is prepped and ready for consumption by model creation, this step will focus on identifying clusters of the product lines based on the selling pattern.

As similar categorization is not available earlier, unsupervised K means model will be used. Though heuristic, this model will provide additional advantage of speed when dealing with 8 years' worth of data across all the product lines and its sale entries.

As a pre-requisite to establishing the "k" value, we will run the validation set for multiple values of "k" and use elbow chart to find the optimal k value to determine the number of clusters accounting into Business' logical groups.

*Given the prepped data, use K means to group over 10k product lines to generate new feature of Product groups which will be used for Time series model.

Step5: ARIMA

As we have historical data available for each product/product line and trends are clear, we will use the time series analysis approach to demand forecasting. We will account into seasonal fluctuations in demand, cyclical patterns, and key Sales trends for each pattern with information from the "External" data set.

ARIMA model can be setup without exogenous regressors and by including weather and availability of raw materials in the model to lead to the model to a better R squared and Mean Absolute Percentage Error at a warehouse level.

*Given the cluster from K-means, use time series analysis to forecast the 12- month demand model for the product groups

Step6: Regression (Data Refresh/ Model tuning)

As demand forecast will be an ongoing process of testing and learning that involves driving an intelligent and agile response to demand and reduce bias and error over time, we will constantly tune the model with the new set of parameters, say every month/quarter agreed by business. For this scenario, we will establish a baseline regression model to understand the relationship of demand vs. the regressors accounting the current scenario. As a constant learner, the model will account the newly established regression relationship to account the new criteria in the regression model

- Given the baseline regression model based on output from cluster and Arima models, use Regression to identify up-to-date relationship and incorporate the changes to the predictors' interactivity with response to maintain the model agility.

Best Practices:

Code Reuse: As we have created a standardized way of clustering products, it is a good practice to establish a data store with the new feature in conjunction with the product lines to reuse the output of the model for ongoing/future initiatives with the assigned clusters. Code re-use will always help cut work time down by a measurable amount.