**IDS 561 Analytics for Big Data -**

**Final Project Report**

**Group – 13 Inventory Demand Prediction**

**Team Members:**

**Aruna Venkatasubramaniam 678860489**

**Gughanraj Selvaraju 662075168**

**Meenakshi Janardanan 671917004**

**Problem Setting:**

A national fresh bakery in Mexico wants to estimate their weekly consumer demand. If they produce too much baked goods and there is no enough demand, the products will get returned/expired and if they don’t produce enough goods, revenue won’t be maximized. Therefore, accurately estimating demand is important. Currently, daily inventory calculations are performed by direct delivery sales employees who must single-handedly predict the supply and demand based on their personal experiences with each store. With some breads carrying a one-week shelf life, the acceptable margin for error is small. In this project, we are going to develop a model to accurately predict the inventory demand based on historical sales data.

**Goal:**

The goal of the project is to maximize sales and minimize returns of the bakery goods. In order to do that, we are going to analyze and predict the demand of a product for a given week, at a store.

Our input is the sales transactions of these bakery goods for 9 weeks and we will forecast the demand for a given week. Each transaction consists of sales and returns. Returns are the products that are unsold and expired. The demand for a product in a certain week is defined as the sales this week subtracted by the return next week.

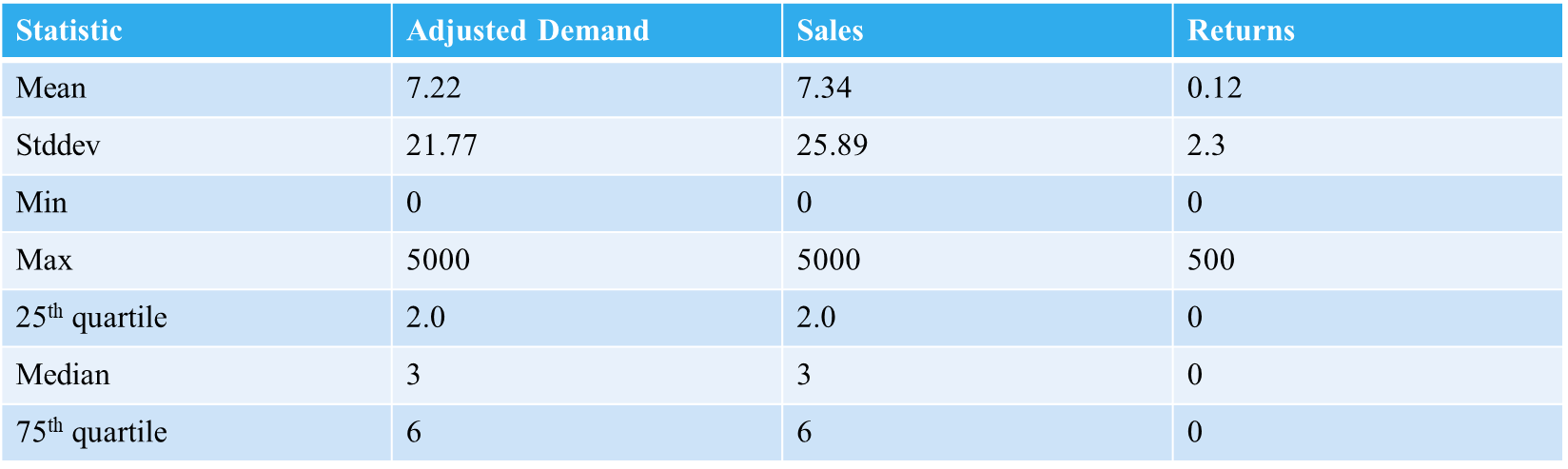
**Data source and Variables:**

The data is obtained from Kaggle - <https://www.kaggle.com/c/grupo-bimbo-inventory-demand/data>. The variables in the dataset are as follows:

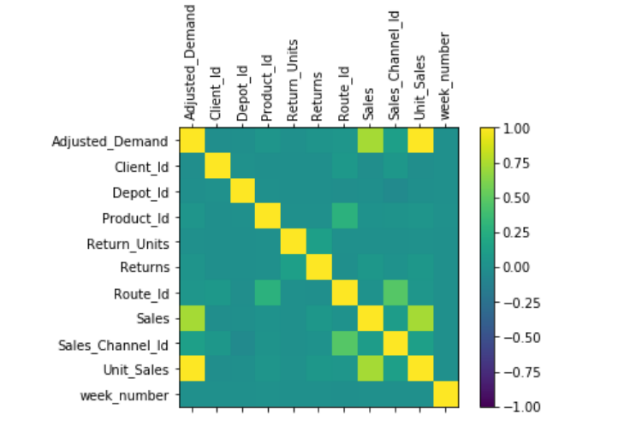


**Descriptive Statistics**

* The demand for each client product pair has a minimum of 0 and a maximum of 4777. The majority of the demand lies between 0 and 6, which means the data has a lot of variance.
* Same goes for Returns all the returns are 0 but the maximum value is up to 500.



**Correlation Analysis:**



* Since Adjusted Demand is a function of Sales and Returns, we can see there is high correlation between them.

**Feature Engineering:**

* Extracted the sub brand and product weight from product data. Similarly, combined District and Town from the location data with train data

**Modelling Steps:**

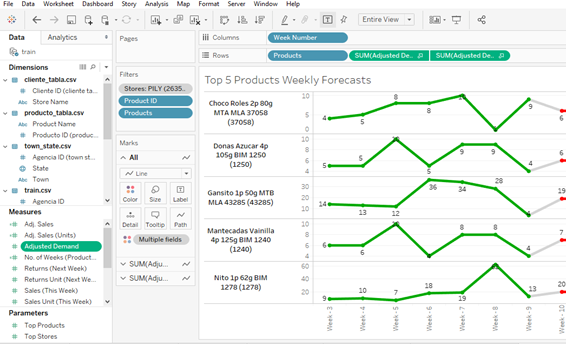
Following steps are followed to create the models in spark:

* Combined new features with original data
* One hot encoding of Categorical variables
* Created Spark modelling pipeline
* Train – Test split of data based on the weeks
* Created models in MLib for train data and Forecast demand for the test data

**Results:**

|  |  |
| --- | --- |
| Techniques | RMSE |
| Random Forest Regressor | 11.827 |
| Gradient Boosted Regressor | 8.513 |

**Tableau Dashboard – Showing the Trend and Forecast:**



**Results & Conclusion:**

Gradient Boosting Regressor achieved the best result with RMSE value of 8.51. We were able to achieve this result using the computational power of spark. Ran the models only for 100 client-product pairs but even that required a lot of computational power and threw memory in a normal CPU setting. Spark with its distributed setup help us efficiently model the data to achieve crucial inferences.

**Role of each team member in the project:**

* Project Selection – Equally
* Setting up the programming environment and Correlation analysis– Gughanraj Selvaraju
* Data Cleaning – Aruna Venkatasubramaniam
* Feature Engineering – Meenaakshi Janardanan
* Pipleine Creation and Modelling & Evaluvation – Gughanraj Selvaraju
* Tableau – Aruna Venkatasubramaniam
* Presentation – Meenaakshi Janardanan
* Final Project - Equally