Sp23 DEMAND/PRICING ANALYTICS

Meal Demand Forecasting

McCombs School of Business
The University of Texas at Austin

Rianna Patel (rnp599)
Rishabh Tiwari (rt27739)
Meeth Yogesh Handa (mh58668)
Parthiv Borgohain (pb25347)
Saurabh Arora (sa55445)

Table of Contents

I. Introduction

The food delivery service is expected to grow to $320 billion by 2029, causing it to potentially have a huge impact on the economy and individual households. This industry uses perishable ingredients and materials at a large scale, making it especially important to have an accurate demand forecasting model to minimize food waste and costs. Given that this service operates based on customer orders, determining the proper inventory to keep on hand at a given time is difficult to estimate. Without the necessary ingredients on hand, the company will be unable to offer the necessary meals or services, leaving customers to transition to competitors in the industry. Not only will this create efficiencies, but it also has benefits in improving supply decisions, optimizing tradeoffs, cost savings, pricing structure, and customer satisfaction. An accurate demand forecasting model will ensure that the company has the right amount of stock on hand, along with sufficient safety stock margins, to ensure it is able to fulfill customer demands with minimal waste. Additionally, with the right amount of stock on hand, the company will be able to find the optimal balance between inventory and demand, as well as create streamlined delivery plans. This in turn will allow for more efficient cost savings with stricter operating cost control and packaging cost savings. Furthermore, with accurate demand forecasting, cash flow predictions will be improved and contribute to more precise product pricing.

Our client is a meal delivery service company that dispatches meals to customers from their fulfillment centers across multiple cities. The goal of this project is to accurately forecast the demand for a meal delivery company in multiple cities to enable their fulfillment centers to plan their raw material stock and staffing needs accordingly. The project aims to develop a reliable demand forecasting model to help the company optimize its inventory management and staffing strategies, while ultimately improving customer satisfaction and profitability.

II. Data

The data used in this project includes 4 datasets: 1) Meal Information, 2) Fulfillment Center Information, 3) Training Data, and 4) Testing Data. The meal information dataset identifies the meal being served to the customer and includes the meal id, category, and cuisine. The Fulfillment Center information dataset includes the center id, city code, region code, center type, and operation area of the center to identify each fulfillment center. To train our models on the data, the training dataset includes various metrics to report the demand for each fulfillment center, such as the number of orders, checkout price, and center id. Lastly, the test set includes the same fields as the training set except the response variable (number of orders) which will be used to provide forecast values for

III. Methodology

A. Exploratory Data Analysis

After merging the datasets to consolidate the necessary data for further modeling, we found some unique insights into customer preferences and business strategies. In **Figure 1** below, we can see that most orders placed were for Italian cuisines, with Thai and Indian following closely behind. This could indicate customer preferences to meals that the client offers, as well as disinclination to other cuisines offered.

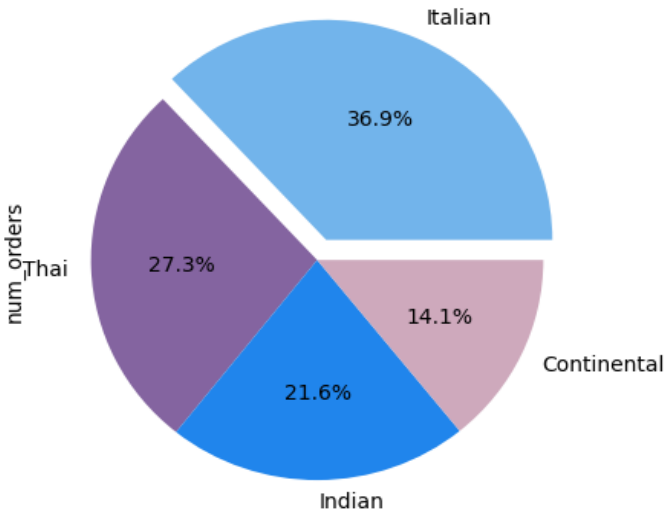

Figure 1: Number of Orders vs. Cuisine

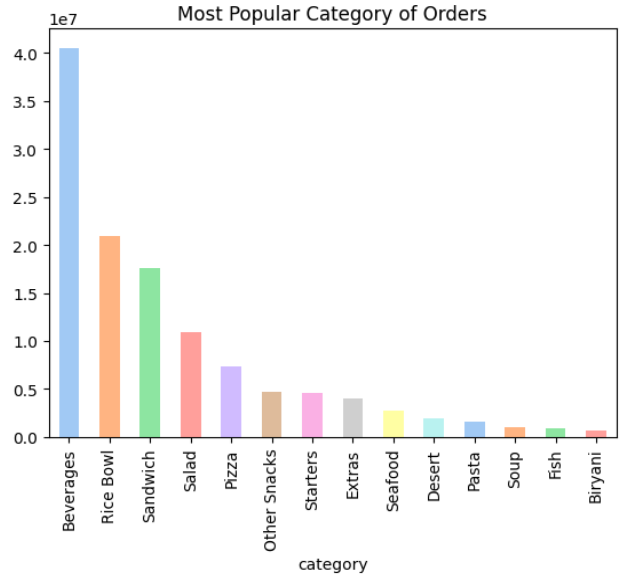

Figure 2: Number of Orders vs. Category

Furthermore, **Figure 2** shows the most popular category of orders with the most popular category as beverages and the least popular as biryani. This shows that the client may have promoted the sale of beverages with coupons or other promotional offers. Assuming that the client continues to promote beverage sales at the scale that they currently do, we can see in this preliminary analysis that the demand for beverages is high and therefore the inventory on hand for these products must be accordingly high in comparison to products such as biryani. As we move further through our analysis we hope to see more specifically how these features relate to

customer demand and how our client can account for these relationships in its business operations.

B. Baseline Models

To contextualize the results of models trained in our analysis and establish baseline performance, 2 baseline models have been created without feature engineering or tuning: 1) a Linear Regression model, and 2) a Decision Tree model. The baseline linear regression model returned a modest performance with 42.29% accuracy as seen below. Furthermore, with a mean squared error of 87598.83, the model demonstrates a large degree of error between the predicted values and real values. **Figure 3** below further shows that the linear regression model is not the best fit for our data given the relationship between predicted y and residual values.

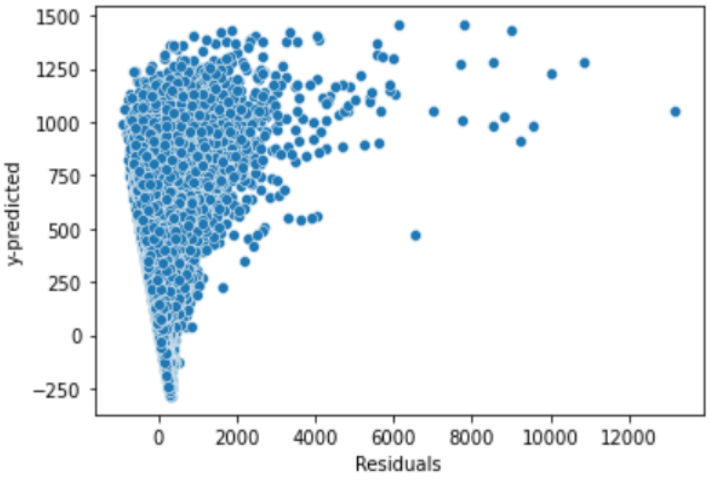

Figure 3: Linear Regression Residual Values

Similarly, the decision tree model produced an accuracy of 69.18% and a mean squared error of 46784.9, using a max_depth of 10 and min_samples_leaf of 8. This model is a better fit for our data than the linear regression model, indicating that our data is highly non-linear and demonstrates a complex relationship between the dependent and independent variables. While these models are not considered highly accurate, they provide a starting point for further analysis and refinement of our models.

C. Data Preprocessing

We prepared the data for modeling by splitting the data into a training and validation set with a 80/20 split. Additionally, we scaled the numeric features to

minimize the distance between features and performed one-hot encoding to convert categorical features, such as city code, category, and cuisine, into a format that can be processed in our models. Upon further analysis of our features, we found that the checkout price and base price features are highly correlated as seen in **Figure 4** to the right.

This intuitively makes sense since the checkout price is likely to be a sum of the base price and various fees, such as taxes or delivery fees. Due to this highly correlated relationship, we decided to remove the checkout price from our analysis without significant information loss.

III. Results

To forecast our client's future meal demand, we created five models: 1) Linear Regression (Simple and Dynamic), 2) Decision Tree, 3) Random Forest, 4) Light GBM, and 5)XGBoost. With these models, we are able to predict the weekly number of orders expected for the client's centers.

A. Linear Regression

a) Simple Linear Regression

Linear Regression poses many advantages for finding relationships in linear data, however, may struggle with finding more complex relationships in our dataset. While this method is easier to implement a train computationally, it tends to be prone to overfitting and noise. Initially, the Linear Regression model returned an R-Squared score of 43.11% on the validation dataset, slightly higher than the baseline Linear Regression model. However, after adding Ridge and Lasso penalties, we found that the Ridge Regression model with the parameters below produced the highest R-Squared accuracy rate of 43.2% and an MSE of 89,397.35. {'alpha': 1.0, 'fit_intercept': False}. In this hyperparameter set, "alpha = 1.0" specifies the constant that multiplies the L2 term and controls regularization strength. Additionally, the final hyperparameter specifies that no intercept is used in this regression to fit the model.

b) Dynamic Demand Linear Regression

To further test our linear regression model we stored the average of demands for every food item in every center (for imputation purposes if the past weeks data does not exist) and ran a **regression model using the lagged demand features**. After tuning the hyperparameters to the same that were included in the previous Linear Regression model, our new Ridge Regression

model returned a R-Squared score of 57.91% and an MSE of 68,735.3. Furthermore, the lagged demand feature has the 6th highest coefficient of 200.12, indicating it has a strong effect on the forecasted demand. We can see that the inclusion of lagged demand has significantly improved the accuracy of the Linear Regression model. This helps us build a dynamic demand model for our client which would explain current behavior of demand trends as well as forecast future values.

B. Decision Trees

After hyperparameter tuning for the Decision Tree model, the best hyperparameters found were {'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 10}. This combination produced a R-Squared score of 75.02% and an MSE of 37,915.51. We can see that this model is significantly more accurate than the Linear Regression model. Similar to the difference in our baseline models, this could be due to the fact that our data is highly non-linear and has a complex relationship between the number of orders and predictor variables.

C. Random Forest

Random Forest regression is similar to the previous model in the way that it builds a "forest" with an ensemble of decision trees. Unlike the Linear Regression model, Random Forest works well with both non-linear relationships within our data and is not highly influenced by outliers. After hyperparameter tuning, we found the following hyperparameters to produce the best fit for our data: {'max_depth': None, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 100}. As a result, the Random Forest model performed better than both the Linear Regression and Decision Tree model with an R-Squared of 80.61% and an MSE of 29,420. Furthermore, we can see in **Figure 5** below which figures have the highest influence on the number of orders used in this model. The most important feature shown is the base price, indicating that what influences customer orders the most is the price of their meal, excluding extra fees.

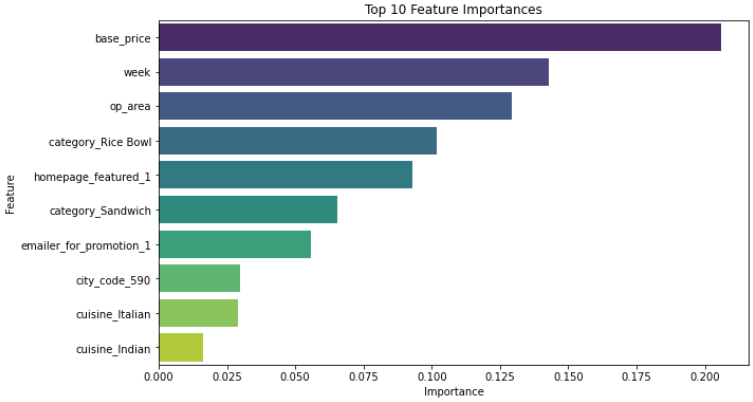

Figure 5: Top 10 Features - Random Forest

D. Light GBM

Light GBM is a gradient-boosting model that also uses tree-based algorithms to build the model. A main advantage of this model is that it tends to produce a higher accuracy since it follows a leaf-wise split approach rather than level-wise splits. However, this can cause the model to be prone to overfitting when testing new data. To find the best model for our data, we used the hyperparameters {'learning_rate': 0.2, 'max_depth': -1, 'min_child_samples': 10, 'num_leaves': 100, 'reg_alpha': 0}. This resulted in a model with an 82.15% R-Squared and a 27,085.49 MSE. **Figure 6** below shows the top 10 feature importances from this model. We can see that these are very similar to those from the Random Forest model, however, the Week feature is ranked first, above the Base Price.

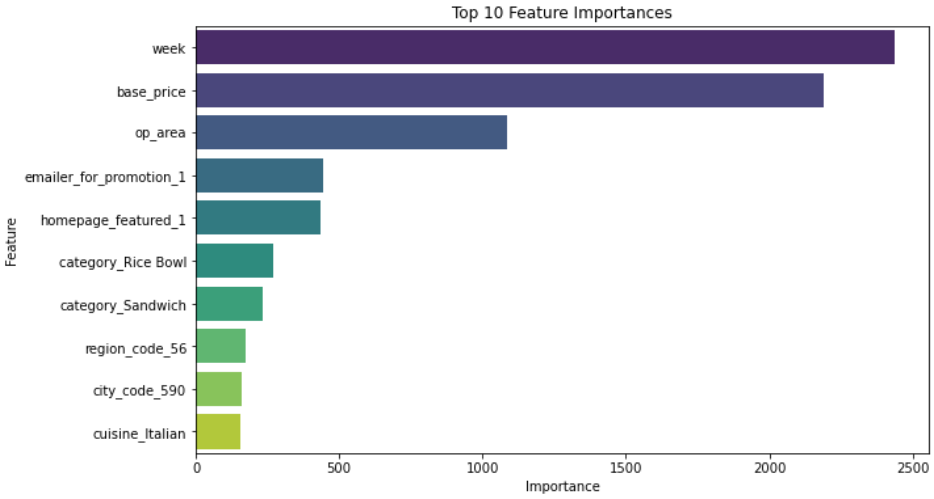

Figure 6: Top 10 Features - Light GBM

E. XGBoost

Lastly, XGBoost is another gradient-boosting model that uses tree-based algorithms but is slightly slower than the Light GBM algorithm. The hyperparameter tuning for this model found the following hyperparameters to produce the highest accuracy rate: {'min_child_weight': 7, 'max_depth': 15, 'learning_rate': 0.1, 'gamma': 0.3, 'colsample_bytree': 0.7}. As a result, the model returned an R-Squared score of 84.18% and an MSE of 24,007.14. This model performed the best compared to the previous models discussed according to these metrics. We can see the most important features used in this model in **Figure 7** below.

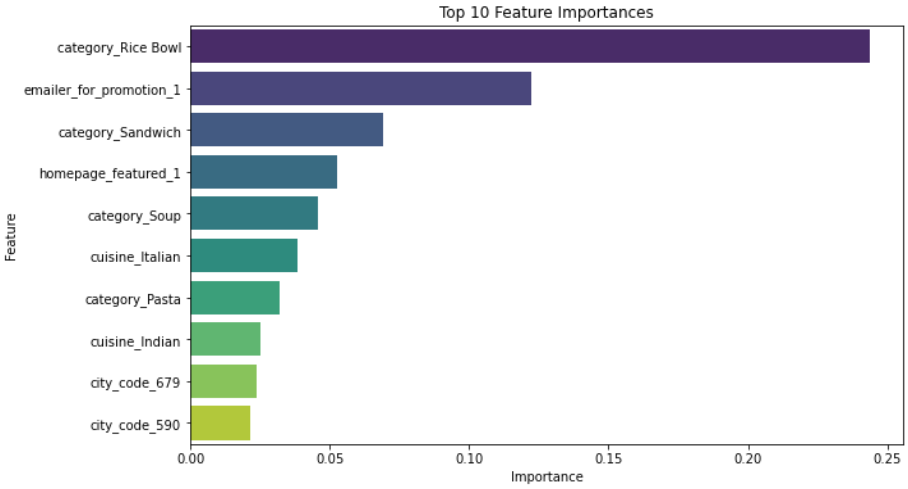

Figure 7: Top 10 Features - XGBoost

IV. Conclusion

After analyzing the results from each of our models with the unique hyperparameters, we determined that the XGBoost model performed the best on the dataset to forecast the meal delivery demand with an R-Squared of 84.18% on the validation set. The feature importance results provided with different models serve as key indicators of future demand.

The output of the demand forecast has been saved and stored in a csv providing easy access to business stakeholders. With the forecasted demand expected from the XGBoost model, we expect that our client can utilize these results to improve 4 major areas of the business:

1) Inventory Management - the meal delivery company will be able to stock an accurate level of raw materials, thus reducing potential waste

2) Procurement Planning - based on accurate demand forecasting, the company can source materials, labor, and vendors to produce quality meals

3) Staffing Decisions - with the expected demand, the company can make more informed staffing decisions to ensure the centers minimize under and over-staffing

4) Customer Satisfaction - by reducing out-of-stock meals and timely deliveries, the client can further increase customer satisfaction