

**Term Project Proposal (Milestone 3 Updated with Preliminary Analysis)**

David Berberena, Brian Mann

Bellevue University

DSC 630 Predictive Analytics

Andrew Hua

July 14, 2024

## **Term Project Proposal (Milestone 3 Updated with Preliminary Analysis)**

### **Introduction**

The restaurant business is one of those fast-paced, ever-evolving industries that either helps investors ride a wave of success or drowns them in the depths of missed opportunity. Roughly sixty percent of new restaurants close their doors within the first three years of operation. Owners, operators, and stakeholders must consider many factors when opening and running a restaurant, with one of the most important being customer analytics. This allows the business to deliver the best experience for the customer, thereby raising the chance of success (Letchinger, 2013).

When looking at companies that have built multiple large-scale restaurant brands, their shareholders need assurance that the company knows what their customers want so that revenue continues to increase. Yet here lies a major business problem: how can large restaurant conglomerates understand and cater to their customers' needs when corporate suits never fraternize with the day-to-day guests?

Take Bloomin' Brands for example, or maybe Yum! Brands, or even Darden Restaurants. With so many restaurant brands under their belts with different cuisines and concepts, how can these conglomerates predict which sector will be the most lucrative? Data holds the answer to this problem, as these restaurant conglomerates constantly intake data and process it to make certain that revenue is maximized. One thing that shareholders would be interested in is how data can predict the revenue of various sectors of the company. Being aware of what their customers want ultimately leads to more informed decision-making, resulting in a larger net profit for the company and shareholders (Smilansky, 2023).

## The Data

To participate in the predictive analytic process of restaurant conglomerates, we have accessed a Kaggle dataset of simulated data outlining specific factors that affect monthly restaurant revenue (Abdurakhimov, 2024). This dataset provides the following variables: customer count, average menu price, marketing expenditures, cuisine type, average customer spending, the presence of a restaurant promotion, review count, and monthly revenue. As restaurant revenue is the target variable in this proposal, the other variables will act as predictors to maximize future restaurant revenue growth. There are aspects of the data that we will need to further clarify, as most of the variables within our dataset do not contain explicit units of measure. For example, the number of customers variable does not specify in what time period this observation is valid. As this data is simulated, we shall place our own units of measure as defined below:

**Number of Customers:** This variable will be the average hourly customer count for the month

**Menu Price:** Here will lie the average menu price of all entrée items the restaurant sells

**Marketing Spending:** This variable will be represented in thousands of dollars

**Average Customer Spending:** Here will define the average total check amount per customer

**Promotion:** We will establish this variable as the presence of a Happy Hour promotion or similar in-store promotion, such as a monthly coupon redeemable upon dine-in.

Considering these new assumptions given to the dataset, we cannot guarantee that the results accurately reflect a real-world scenario. This is due to the nature of the data being simulated and the missing meanings behind the original dataset.

### **Model Selection and Performance Metrics**

Once our data is cleaned and processed, we will need to transform any categorical variables into numerical variables via one-hot encoding and/or dummy variable generation. The dataset will then be split into training and test sets at an 80/20 ratio. Once these transformations are made, it will then be ready for modeling.

We will be opting to use linear regression as the primary means of predictive modeling. Our choice of linear regression to verify the factors that positively affect restaurant revenue will allow us to see which factors should be presented to shareholders of restaurant conglomerates like Darden Restaurants as growth opportunities. Providing a promising model will create a secure space for informed decision-making, which could lead to increased investments in more lucrative sectors and potential rebranding efforts for struggling restaurants.

We will be employing Jupyter Notebook to craft a multiple linear regression model in Python, with the goal being to identify which factors are most influential in generating higher monthly restaurant revenue. This model assumes that the predictor variables are linearly related to the outcome variable. To account for potential biases within our models, we will use k-fold cross-validation techniques and L1/L2 normalization (Lasso and Ridge regression respectively) as part of the hyperparameter tuning process. The goal of the tuning process is to minimize the root mean squared error (RMSE) value and maximize the coefficient of

determination ( $R^2$ ) so that a trained model can accurately predict monthly revenue. Minimizing RMSE limits the chance of inaccuracies while maximizing  $R^2$  demonstrates what proportion of the variance can be predicted by the model (Chugh, 2020).

## **Learning Goals**

During this project, we will consider the following questions:

- What factors have the most impact on restaurant performance?
- Can restaurant revenue be accurately predicted based on the data we have been given?
- Are marketing promotions worth conducting for the company?
- How much should each restaurant be spending on marketing to maximize revenue?
- What did modeling help explain about the data that exploratory data analysis could not?

These questions will guide us during each phase of the learning process. By the end of the project, we hope that the answers to these questions will ultimately lead to more informed decision-making from shareholders.

## **Risks and Ethical Implications**

Several risks must be considered during the course of the project. One is overfitting. Since the dataset we are working with is not particularly large (approximately one thousand rows of data), we should take care to avoid overfitting our models to the data. Similarly, we must also try to avoid multicollinearity. It may be the case that the number of customers, average spending, reviews, and other factors have a high degree of correlation with one another. Another risk is that the data we are using is machine-generated. The data might not

necessarily be based on a real-world company, thus making it difficult to make meaningful generalizations about the restaurant industry. Without taking steps to minimize these risks, we increase the chances of inaccurate predictions, overly complex models, and an inability to generalize across different datasets (Lindgren, 2019).

Ethically, we must simply make it clear that this data is not meant for making broad generalizations about the restaurant industry. We have made several assumptions about the dataset and what it represents. These may not be wholly accurate, and we should be careful so as not to misrepresent what we have stated as fact. Consequently, we will not be using this data for any sort of financial gain, as this is for educational purposes only.

### **Contingency Plan**

While we believe that our proposal is efficient and considers many aspects of the business problem and how the data we have acquired can provide an adequate answer, the model and its performance metrics may prove to be ineffective. Should the model not be viable for deployment from our first attempt at streamlining the initial model, there are a few scenarios that we can address as a plan to combat the unexpected.

- **Underfitting:** If our model proves to be underfit for effective analysis, we may need to seek out additional data, or another data source altogether. One such data source is the restaurant revenue prediction dataset from a Kaggle competition in 2015 (Ozer et al, 2015).

- **Overfitting:** If the model has been overfit as shown by inaccurate metrics, we should adjust hyperparameters or seek alternative models such as random forest regression or K-nearest-neighbors regression.
- **Variability:** If there is too much variability in the data, there may be confounding factors that should be taken care of. We may need to then remove outliers, mathematically transform certain variables, or get rid of some variables altogether.

### Milestone 3

#### Initial Takeaways

After performing our initial exploratory data analysis on our primary dataset, we have realized that we will not need our contingency dataset, as the information gleaned from the EDA can help answer each of the questions we have formulated about the data. Looking back on our previously outlined learning goals, the primary dataset shows the features affecting monthly restaurant revenue. Through the calculation of variable correlation in our preliminary analysis and our intention of finding which factors are most significant in affecting monthly restaurant revenue using a linear regression model, we can move forward in the pursuit of the answer to identify the most impactful factors on restaurant performance. As our dataset contains relevant features for the task of predicting monthly restaurant revenue, we are confident that the data can help us make accurate predictions. The next question we will address is how accurate those predictions will be. The answer will come with the generation of a linear regression model and the calculation of the relevant accuracy metrics (R-squared, mean absolute error (MAE) and root mean squared error (RSME)).

To answer our initial question of whether marketing promotions are worth conducting, we made sure to explore this within our EDA by grouping our data by the presence of a promotional campaign, computing the average monthly revenue, then comparing the two figures. It was made clear that there was only a one percent difference in revenue in favor of a restaurant having a promotion, so we can determine that it is not worth the effort for restaurants to try and entice customers through a promotional campaign. Regarding how much a restaurant should spend on marketing to maximize their monthly revenue, our initial analysis has revealed that there is not a surefire way to answer this question based on the current data. The reasoning behind this is due to the data not being exponentially distributed to see at what point in marketing expenditure the increase in monthly revenue begins to plateau. After crafting visualizations that confirmed this, it was evident that the data follows a linear pattern of distribution, pointing us back in the direction of linear regression for our model analysis. We are on track to address the final question in the model generation stage of the analysis of the data.

## **Visualizations**

In our EDA, there were key visualizations that aided in answering a few of the above questions and helped to tell the story of our data. Bar charts were used to compare aggregated average monthly restaurant revenue by restaurant cuisine types (i.e. Japanese, American, Mexican, Italian) as well as which of these cuisine types was most prevalent in the dataset. The creation of a heatmap aided in finding the most highly correlated features to monthly revenue within the dataset, allowing us to keep our eyes on those features when the model generation



and evaluation stage occurs to confirm whether these features are significant as well as linearly related. Histograms of each of the numerical variables indicated that they were all uniformly distributed apart from monthly revenue, which was normally distributed. The last visualization that allowed us to directly answer one of our previous questions was a scatter plot outfitted with a line of best fit to see the distribution of data for monthly revenue against marketing expenditure.

To craft these visualizations, we adjusted the data as stated previously by transforming the categorical variables within the dataset to numerical variables via dummy variable construction. After verifying that no missing values were present and each column's data type was correct, the visualization creation process was straightforward. After seeing the scatter plot graphic attached to marketing expenditure, we believe that it would be best to divert our attention away from this path of exploration as the data does not support a potential answer as to what dollar amount would maximize restaurant revenue for the month. The data is ill-equipped for such an inquiry, and more observations would need to be gathered.

### **Model and Expectation Confirmation**

Now that we have undergone our EDA process, we are confident that linear regression is the correct model choice for the questions we have asked of the data. The specific model that we will employ will be the ordinary least squares model of linear regression, as this model provides a high-level summary of the model's results for easy discernment of key metrics and deductions. This model also allows us to calculate another performance metric that is crucial to the identification of the most impactful factors on monthly restaurant revenue: the p-value

statistic. This figure will allow us to determine which features have the most significance in affecting monthly revenue, directly answering the first of our few dataset inquiries. This addition to the set of metrics we will be evaluating will make certain that the performance of the model is verified and the predictors are classified as either impactful or irrelevant, and to what degree these factors might be categorized as such.

Our expectations of the project have been solidified as appropriate for the dataset we have obtained based on the results of our initial analysis. We have checked our visualizations to make sure that they not only are relevant to the business objective, but also guide us to that end with insightful meaning upon gleaning understanding from them. We have determined that the variables marketing expenditure, menu price and total customers have a slightly greater correlation with revenue, and we will make sure to take this into consideration during the model building phase. With our data poised for ordinary least squares linear regression model generation and evaluation, we are eager to see what the model presents in its predictive power for monthly restaurant revenue.

## References

Letchinger, C. (2013). *The Anatomy of Restaurant Failure: Dead Man Walking*. Menu Cover Depot. <https://www.menucoverdepot.com/resource-center/articles/restaurant-failure/>

Abdurakhimov, M. (2024). *Restaurants Revenue Prediction [Data set]*. Kaggle. <https://doi.org/10.34740/KAGGLE/DSV/7420974>

Ozer, E., O'Connell, M., Kan, W. (2015). *Restaurant Revenue Prediction*. Kaggle. <https://kaggle.com/competitions/restaurant-revenue-prediction>

Smilansky, V. (2023). *Data-driven decision making: How to use data to make more informed decisions*. ThoughtSpot. <https://www.thoughtspot.com/data-trends/best-practices/data-driven-decision-making>

Chugh, A. (2020). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?*. Medium. <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>

Lindgren, I. (2019). *The Dangers of Under-fitting and Over-fitting*. Medium. <https://medium.com/analytics-vidhya/the-dangers-of-under-fitting-and-over-fitting-495f9efa1847>