Predicting Monthly Revenue in the Restaurant Industry

(Milestone 4 Updated with Finalized Results)

David Berberena, Brian Mann

Bellevue University

DSC 630 Predictive Analytics

Andrew Hua

July 14, 2024

Predicting Monthly Revenue in the Restaurant Industry

(Milestone 4 Updated with Finalized Results)

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.

Milestone 4

Data Preparation

For our dataset to be ready for the ordinary least squares model's creation, one of the variables (Cuisine Type) needed to be converted from categorical to numerical using Pandas's get_dummies() function. Then the data was split into the appropriate training and test sets at an 80/20 ratio for the model to learn enough about the data to make predictions on monthly restaurant revenue. The goal initially was to create a pipeline that allowed the creation of the ordinary least squares model after the data was standardized using the StandardScaler() function, yet the OLS model is not compatible with a pipeline as is.

To alleviate this concern, a custom wrapper that held an OLS model within itself was created so the model would work with the pipeline. A search space was then crafted to account for the new OLS wrapper and Ridge Regression, which also deals with any potential bias present in the OLS model. With the OLS wrapper and the search space ready for model generation, a grid search was conducted using GridSearchCV() and was customized to cross-validate five times using both the pipeline and search space to output each OLS model summary and the best pipeline estimator along with its matching R-squared and RMSE statistics. The OLS model was now ready to be trained and fit to the data.

Model Generation and Results

The best-performing OLS model summary that came from the grid search is as follows:

OLS Regression Results

Dep. Variable:	Monthly_Revenue	R-squared:	0.692
Model:	OLS	Adj. R-squared:	0.687
Method:	Least Squares	F-statistic:	156.9
Date:	Wed, 24 Jul 2024	Prob (F-statistic):	1.76e-154
Time:	10:00:49	Log-Likelihood:	-3501.0
No. Observations:	640	AIC:	7022.
Df Residuals:	630	BIC:	7067.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	272.0739	2.290	118.819	0.000	267.577	276.571
x1	75.8445	2.301	32.956	0.000	71.325	80.364
x2	25.5013	2.300	11.086	0.000	20.984	30.018
x3	27.4173	2.312	11.861	0.000	22.878	31.957
x4	-0.8585	2.298	-0.374	0.709	-5.371	3.654
x5	-3.6514	2.302	-1.586	0.113	-8.171	0.869
x6	-1.0937	2.301	-0.475	0.635	-5.612	3.424
x7	-0.9335	2.765	-0.338	0.736	-6.364	4.497
x8	-2.6698	2.808	-0.951	0.342	-8.184	2.844
x9	-0.9220	2.802	-0.329	0.742	-6.425	4.581

Omnibus:	2.153	Durbin-Watson:	1.816
Prob(Omnibus):	0.341	Jarque-Bera (JB):	2.218
Skew:	-0.137	Prob(JB):	0.330
Kurtosis:	2.909	Cond. No.	2.04

As seen above, the R-Squared value is 0.692, meaning that 69% of the variance within the data can be explained by the model. After running the grid search, it was found that the best estimator of the data was a Ridge regression model with the R-Squared value increasing very negligibly, as shown by the code snippet below:

```
Best estimator: Pipeline(steps=[('scale', StandardScaler()),
                ('model', Ridge(alpha=7.9060432109076855, max_iter=10000))])
R^2: 0.6937014067272136
RMSE: 60.856889064989275

Feature importance:
Number_of_Customers: 74.5995
Menu_Price: 24.5487
Marketing_Spend: 26.0052
Average_Customer_Spending: -0.5611
Promotions: -3.5897
Reviews: -1.3689
Cuisine_Type_Italian: 0.4578
Cuisine_Type_Japanese: -0.6830
Cuisine_Type_Mexican: 0.2322
```

The RMSE statistic for the best estimator model is 60.85, which seems to be a small value with respect to the concept of monthly restaurant revenue, which is typically measured in thousands

or higher. However, since the dataset's revenue values range in the hundreds, this RMSE value indicates a large amount of error in the Ridge regression model's predictions against the test set observations. Looking at the OLS summary above, the Prob (F-statistic) value is extremely small, indicating that the model is statistically significant and that the features are collectively impactful as predictors for monthly restaurant revenue. With this being said, the Ridge regression model defined as the best estimator exhibits high statistical significance to the data, a high amount of predictive error, and the ability to explain almost 70% of the data's variance after adjusting for potential bias.

While our primary model choice has been properly created and cross-validated, other model choices were also explored. A decision tree regressor was crafted to see if the evaluation metrics would increase in performance with this model choice as decision trees can handle both classification and regression issues, yet the following output shows that the decision tree model performed worse than the Ridge regression model.

```
Best estimator: DecisionTreeRegressor(max_depth=10, min_samples_leaf=4, min_samples_split=10,
                      random_state=1)
RMSE: 77.33858179980125
R^2 Value: 0.5053273330434154
```

With the R-Squared value at 0.5, only 50% of the data's variance is explained by the decision tree as opposed to the 69% justified by the Ridge regression model. The amount of error between predictions and test set observations is higher with the decision tree model as well, with the RMSE at 77.33 while the Ridge regression model's RMSE stands at 60.85. Then an ensemble model was crafted using a stacking regressor model encompassing Ridge regression, a decision tree model, a random forest model, and a gradient boosting algorithm. This conglomerate model, while performing better than the decision tree model, did not eclipse the performance of the Ridge regression model. After studying the

performance metrics of the decision tree and ensemble models, it is clear that the OLS model subject to L2 (Ridge) regularization provided the best performance for the data.

Data Analysis Recommendations

After the full analysis of the data with exploratory data analysis, data visualization, and model generation, the initial learning goals that were not addressed in full prior to this point can now be fully vetted. To answer the inquiry of which factors have the most impact on restaurant performance, the OLS summary shows that the first three variables within the dataset are statistically impactful to monthly restaurant revenue, and those three variables are Number_of_Customers, Menu_Price, and Marketing_Spend. The next question listed on our learning goals was whether restaurant revenue can be accurately predicted based on the data we have been given.

After interpreting the results of the best estimator Ridge regression regularization of the OLS model, it can be determined that the dataset alone does not allow for the accurate prediction of monthly restaurant revenue. Even though the model fits to the data well and can explain a passable amount of variance, the margin of error between predictions and actual observations outlined by the RMSE is too high to consider the model accurate in its predictive power. However, a recommendation that can be made to help increase monthly restaurant revenue based on the OLS model summary and the exploratory data analysis is to focus on marketing expenditure and the number of customers entering the restaurant. Since these factors have been considered impactful by the model with the heatmap confirming these variables to be positively correlated to monthly restaurant revenue, it is important to reexamine how these variables are handled to potentially increase monthly sales. For a

predictive model to provide a higher level of statistical significance and sway for a recommendation, however, more data will need to be gathered.

To better collect more data, variables such as customer demographics, seasonal trends, competitor locations, time of day, method of promotion and other such factors might offer further insights into what is impacting revenue. It could also help to create and distribute customer surveys with questions about their experience, what foods they enjoy most, and how they decided to eat at that particular establishment. Since it was clear from our exploratory analysis that the promotions were having little to no effect, this would help provide insights to fix what went wrong in future campaigns. Additionally, implementing more advanced modeling techniques that improve upon the initial ensemble methods or machine learning algorithms may improve predictive accuracy and provide a more nuanced understanding of the relationships within the data. By addressing these aspects, the model could evolve to offer more precise recommendations and actionable strategies for increasing 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