Predicting Restaurant Revenue with Various Machine Learning Models

David J. Roh*
Department of Computer Science
Texas A&M University
david.roh@tamu.edu

Abstract

In this paper, I employ an array of machine learning models to predict annual restaurant revenues using an obfuscated dataset provided by TFI. This dataset includes factors such as location, opening date, and demographic, real estate, and commercial data. Through exploratory data analysis and feature engineering, I aim to uncover significant correlations and features to enhance the accuracy of my predictions. This paper contributes to understanding the effectiveness of diverse regression models in predicting outcomes in complex and obscured scenarios, offering insights for strategic decision-making in the restaurant industry.

1 Introduction

In the hyper-competitive landscape of the restaurant industry, the margin between success and failure often hinges on a business's ability to accurately understand and predict its revenue. This understanding underpins strategic decisions around operations, marketing, and investments, and can thus significantly influence a restaurant's potential for growth and sustainability. However, predicting restaurant revenue is far from straightforward due to the confluence of numerous factors such as demographics, location, real estate, and the commercial environment. This complexity, combined with the often slim profit margins typical to the industry, makes effective financial planning challenging.

This project is motivated by the call for accurate, data-driven approaches to revenue prediction in the restaurant industry, an aspect that has gained even more significance in the wake of the recent pandemic. As restaurants grapple with evolving customer preferences and a rapidly changing market environment, leveraging data to make informed decisions has become not just an advantage, but a requirement for survival (National Restaurant Association, 2023).

In this paper, I strive towards developing a robust machine learning model capable of predicting restaurant revenue based on a wide array of features, using a dataset provided by TFI. By exploring multiple machine learning methods and conducting exploratory data analysis and feature engineering, I aim to find significant correlations and features that can lead to accurate revenue predictions and insights into what drives restaurant success. In practice and industry, this new information can be used towards optimizing resource allocation.

2 Literature Review

A key aspect of this project involves understanding and building upon the current state of the art in restaurant revenue prediction. My research has led me to explore various sources, including academic literature and Kaggle competitions.

*CSCE 421-200: Machine Learning (Honors), Spring 2023, Texas A&M University, TX. (LaTeX template borrowed from NeurIPS 2022)

Kaggle previously hosted a competition utilizing the dataset I am using for this study (Kaggle, 2015). This competition included features such as the restaurant's opening date, location, type, and various obfuscated demographic data. A number of participants achieved a reasonable degree of accuracy employing models like Random Forest and Gradient Boosting. However, the public leaderboard's accuracy is somewhat concealed, and the methodologies of the top-performing models are not publicly accessible. Despite these limitations, examining the submissions, discussions, and models with available code provided valuable insights and inspiration for this study. I have incorporated some of these insights into my research while developing my unique approach to feature engineering, model selection, parameter tuning, and stacking methods. (Taruto, 2020)

Additionally, I drew insights from academic literature, such as a study conducted by Stanford students to predict restaurant success based on Yelp sentiment analysis and various other characteristics. Their study, which utilized Neural Networks and Support Vector Regression (SVR) as predictive models, yielded interesting findings on the effectiveness of these methods in the domain of restaurant success prediction. They found that using solely restaurant characteristics was difficult to find a strong model that predicted restaurant success (which in their case was star rating). (Kang & Vo (2016))

While these sources have significantly informed my understanding of the problem space and the potential approaches to addressing it, there remains a clear gap in the literature regarding comprehensive and transparent methodologies for predicting restaurant revenue using diverse machine learning models. This paper aims to address this gap by developing and thoroughly documenting a robust machine learning model for restaurant revenue prediction, thereby contributing to the body of knowledge in this area. Even if the model does not achieve a high degree of accuracy, the insights gained from this study can be used to inform future research and development in this area.

3 Problem Formulation

I aim to predict the annual revenue of restaurants based on various given factors. These factors include the opening date of the restaurant, its location, city type, and three categories of obfuscated data: Demographic, Real Estate, and Commercial data. The challenge is to accurately decode and interpret these obfuscated datasets and derive meaningful correlations between them and the restaurant's revenue.

The objective is to develop a (relatively) robust machine learning model that can accurately predict the transformed annual revenue of a restaurant.

This problem is primarily a regression task, since the revenue is continuous. The complexity lies in the high dimensionality of the dataset and the obfuscated nature of some of the features, which makes feature selection, engineering, and interpretation a significant part of the task. Another challenge is the small size of the training set, which makes it difficult to train a model that can generalize well to the test set.

4 Proposed Solution

My approach involves a multi-step process centered around exploratory data analysis (EDA), feature engineering, model selection, hyperparameter tuning, and an ensemble learning method.

4.1 Exploratory Data Analysis and Feature Engineering

Initially, I will conduct a comprehensive exploratory data analysis to understand the data's underlying structure, identify potential patterns, and detect any anomalies or outliers. Next, I will perform feature engineering to create new features that could improve the predictive power of the models, as well as modify existing ones. For instance, the 'Open Date' feature could be transformed into 'Age of the restaurant', which might be a more direct indicator of the revenue.

4.2 Model Selection and Hyperparameter Tuning

My strategy involves the implementation and comparison of several machine learning algorithms, including Catboost, Random Forest, ElasticNet, Gradient Boosting, KNN, Lasso, LightGBM, Linear Regression, Ridge, SVR, Sequential Neural Network, and XGBoost. Each model will be trained

and evaluated on a 80:20 validation split, with hyperparameters, if applicable, modified with GridSearchCV and other Cross-Validation Methods. I also plan to utilize ensemble stacking to combine the strengths of the different models. This technique involves training a meta-model that makes final predictions based on the predictions of the individual models, and potentially improving the overall predictive performance. Since the final test set will be scored based on RMSE, which is sensitive to data that is not normally distributed, I will also explore the use of Log Transformation to normalize.

4.3 Evaluation

The performance of the models will be evaluated using the Root Mean Square Error (RMSE) between the predicted revenue and the actual revenue when submitted to Kaggle. The RMSE is calculated as follows: $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. $R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$. Before submission, I will be using the R^2 and $RMSE$ metrics to evaluate the performance of the models on the validation set.

5 Data Description

The provided dataset from TFI contains 137 restaurants in the training set and 100,000 in the test set. Each restaurant entry includes an opening date, city, city group, restaurant type, and 37 obfuscated features (P1-P37) pertaining to demographic, real estate, and commercial data. The target variable is 'Revenue', indicating a restaurant's transformed annual revenue. This dataset will be used to train, fine-tune, and validate the performance of the predictive models.

6 Results

The results of this study indicate that Gradient Boosting was the best performing model, achieving RMSE scores of 1845784.60 and 1848364.91 on the private test set(99% and 1%), respectively. Random Forest,KNN, and LightGBM also demonstrated reasonable performance, suggesting that with further feature engineering and hyperparameter tuning, these models could potentially yield improved results.

Surprisingly, ensemble stacking, however, did not lead to a significant improvement in performance compared to Gradient Boosting alone, as the chosen models may not have been diverse enough for this technique to be effective. The train and test scores showed a considerable gap, implying potential overfitting in the models, which could have been attributed to the complexity of the models fitting to the noise in the training set rather than towards general patterns.

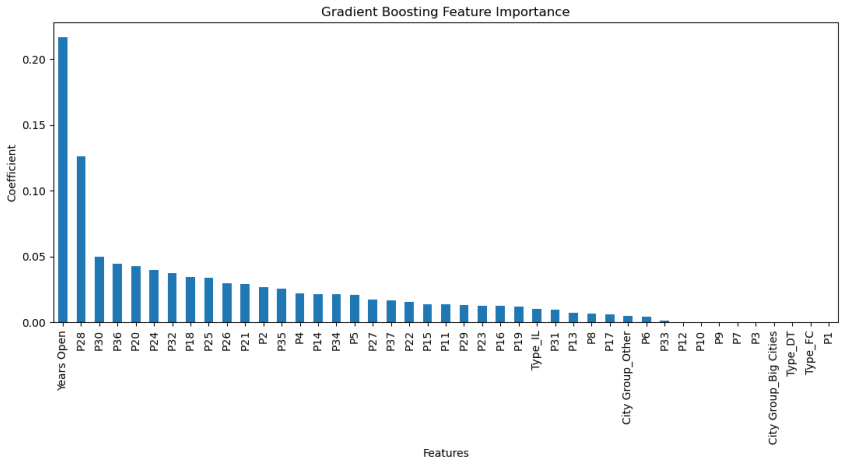

Figure 1: Feature Importance for Gradient Boosting

Parameters for Gradient Boost Found with GridSearchCV

```
Best n_estimators:  20
Best learning_rate:  0.1
Best max_depth:  4
Best subsample:  0.4
Best min_samples_split:  4
Best min_samples_leaf:  5
```

It is clear from Figure 1 that the most important feature that Gradient Boosting finds is most correlated to revenue is how old the establishment is. This makes sense as older restaurants have had more time to build a customer base and establish a reputation.

Regarding feature engineering, the modified datasets with the 'City' column included performed well in terms of RMSE and r^2 metrics on the validation set. However, these models performed poorly when submitted to Kaggle for actual testing. This suggests that incorporating city information may have led to overfitting to the training set, causing decreased performance on the test set.

Balancing between finding good fitting hyperparameters for the training set, while still maintaining generality for the test set was a challenge that involved an element of luck. Unfortunately, the results of this study were not as good as those from other Kaggle notebooks, which may be attributed to my lack of experience in recognizing the best features to use. Gradient Boosting, the best performing model, achieved 65th percentile on the private test set among all submissions.

Model	Private Score	Public Score
Gradient Boosting	1845784.59575	1845597.21043
Random Forest	1848364.9092	1926833.89567
KNN	1875299.94874	1871688.45442
Ensemble Stacking (rf, gb, cat)	1875380.48517	1903724.38684
Ensemble Stacking (gb, rf, knn)	1878737.56525	1893041.42161
LightGBM	1884524.82644	1939470.83391
Ensemble Stacking (rf, cat, xgb)	1896182.15684	1913827.75044
Ensemble Stacking (xgb, cat, rf)	1896924.23344	1913823.67616
CatBoost	1922190.48409	1987687.23056
SVR	1923094.36392	1908418.01027
XGBoost	1992989.75093	1994996.23934
Ridge Regression	2011325.14676	1906020.80884
ElasticNet	3204436.66926	2840187.8722
Lasso Regression	4412700.87596	3833758.81914
CatBoost (Top 3 Cities)	4694827.97527	4515976.4573
CatBoost (All Cities)	4694828.00066	4515976.48783
Sequential Neural Network	14199982.92749	6846907.06139
Linear Regression	112346378.78415	53271479.76005

Table 1: Kaggle RMSE Scores (Descending by Private)

	Train r^2	Val r^2	Test RMSE	Val RMSE
Linear Regression	0.424122	-0.201554	0.349123	0.598522
Lasso Regression	0.108599	0.124444	0.434360	0.510917
Ridge Regression	0.102693	0.059171	0.435796	0.529619
ElasticNet Regression	0.100596	0.104733	0.436305	0.516636
Gradient Boosting Regression	0.532744	0.196290	0.314478	0.489506
Random Forest Regression	0.522611	0.040585	0.317870	0.534825
LightGBM Regression	0.259516	0.117843	0.395887	0.512839
KNN Regression	0.068122	0.004247	0.444112	0.544859
XGBoost Regression	0.369777	0.105839	0.365225	0.516317
Support Vector Regression	0.109819	0.015294	0.434063	0.541828
CatBoost Regression	0.894011	0.138010	0.149777	0.506943
Stacking Regression(rf, cat, xgb)	0.605774	0.095607	0.288859	0.519262
Stacking Regression(xgb, cat, rf)	0.605738	0.095605	0.288872	0.519263
Stacking Regression(rf, gb, cat)	0.740570	0.080570	0.234328	0.523561
Stacking Regression(gb, rf, knn)	0.151420	0.040428	0.423799	0.534868

Table 2: Local Results

7 Limitations and Future Directions

My project could be improved in many areas. Firstly, the feature engineering performed was limited due to time constraints. Although some feature engineering was carried out, such as MICE imputation and log transformation of revenue, more extensive exploration of feature interactions, selection, and transformation techniques could have been conducted. I attempted doing a Pearson correlation analysis but it resulted in mostly weak correlations with a consistent hue across the board and I was not sure how to interpret the results with high information gain.

Furthermore, I could have considered the use of feature selection techniques such as recursive feature elimination to identify the most important features for predicting revenue. Finally, I could have explored the use of feature transformation techniques such as Box-Cox transformation to improve the normality of the data as a whole.

Another limitation was the lack of testing MICE-imputed data in combination with the 'City' feature. Future studies should explore the impact of different combinations of features on model performance, including the use of MICE-imputed data with the 'City' feature across all models but I did this spuriously working with CatBoost.

Furthermore, overfitting appeared to be a concern in the current study, potentially due to the models' complexity. With more time, I could focus on using techniques such as regularization and simpler models.

On top of the models already tested, I could explore the use of more diverse models, such as deep learning techniques and unsupervised learning methods, to enhance the performance of ensemble stacking. I only trained my Sequential Neural Network for 1 hr on both 100000 epochs and with early stopping with not much difference in performance. Possibly I could have tuned the layers to be more general.

Finally, I could consider the removal or transformation of additional features to increase the generality of the models. By focusing on the most important and generalizable features, it may be possible to improve the models' ability to predict revenue for a broader range of restaurants, thus increasing their practical utility in the industry.

8 Conclusion

In conclusion, this study aimed to develop a robust machine learning model for predicting restaurant revenue using a dataset provided by TFI. Through exploratory data analysis, feature engineering, model selection, hyperparameter tuning, and ensemble learning, I attempted to identify significant correlations and features that could lead to accurate revenue predictions. Gradient Boosting emerged as the best-performing model, followed by Random Forest, KNN, and LightGBM, which also showed reasonable performance.

However, there were limitations in the feature engineering performed, and overfitting appeared to be a concern. Additionally, the ensemble stacking method did not significantly improve the overall predictive performance, potentially due to the lack of diversity among the chosen models. Future studies should explore the use of more diverse models and feature engineering techniques to improve the models' ability to predict revenue for a broader range of restaurants.

References

[1] Kaggle. (2015). Restaurant Revenue Prediction. Retrieved from https://www.kaggle.com/c/restaurant-revenue-prediction.

[2] Kang, S., & Vo, V. (2016). Predicting Success of Restaurants in Las Vegas. Retrieved from https://cs229.stanford.edu/proj2016/report/VoKang-PredictingSuccessOfRestaurantsInLasVegas-report.pdf.

[3] National Restaurant Association. (2023). Restaurant Industry Outlook 2023. Retrieved from https://restaurant.org/research-and-media/research/research-reports/state-of-the-industry/.

[4] taruto1215. "Stacking_tutorial_XGB_LGBM_CatBoost_MLP_SVR_KNN." Kaggle, Kaggle, 14 Feb. 2020, https://www.kaggle.com/code/taruto1215/stacking-tutorial-xgb-lgbm-catboost-mlp-svr-knn.