Predicting Restaurant Revenue; Insights from Objective Measurements

Alexia Wells

David Eccles School of Business, University of Utah

MKTG 6620: Machine Learning

Professor Ron Beckstrom

December 15th, 2024

Executive Summary

- Tab Food Investments was interested in predicting the annual revenue of restaurants based on location and other objective measurements.
- An exploratory analysis of the data revealed information about how to proceed.
- Several models were created and evaluated based on their root-mean-square error.
    - Random Forest had the lowest RMSE and is consequently considered the best performing.
- The open date of restaurant locations was a top predictor, but may not be of practical importance.
- The random forest model should be utilized by TFIs development team when determining future restaurant investments.

Introduction

Project Goal

This project aims to develop a model capable of predicting annual restaurant sales based on objective measurements across 100,000 regional locations. Opening new restaurants is time-intensive and costly; selecting a good location is essential to avoid losses. A model that can identify worthwhile restaurant locations would allow TFI to allocate resources in other areas.

Business Problem

TFI is known for expanding restaurant brands internationally through creative marketing and operations (e.g. Popeyes, Arby's, Burger King, Sbarro, etc.). Knowing where to open new restaurants is largely based on the judgment and experience of diverse development teams. Different geographies and cultures can make this a challenging task. When poor locations are chosen, the sites are typically closed within eighteen months. Developing a model to reliably predict a restaurant's success, based on location, is essential for TFI.

The Data

For the purpose of this analysis, two datasets were utilized, train and test. The train set contains 43 total variables and 137 total restaurants/entries. The test set contains 100,000. The columns include the opening date of the restaurant, the location, city type, and 37 other columns from a range of demographic, real estate, and commercial data.

Glimpse of the Dataset

Id	City	City Group	Type	Revenue	Date	P1	...	P37
0	İstanbul	Big Cities	IL	5653753	1999-07-17	4	...	4
1	Ankara	Big Cities	FC	6923131	2008-02-14	4	...	0
2	Diyarbakır	Other	IL	2055379	2013-03-09	2	...	0
3	Tokat	Other	IL	2675511	2012-02-02	6	...	6
4	Gaziantep	Other	IL	4316715	2009-05-09	3	...	3
5	Ankara	Big Cities	FC	5017319	2010-02-12	6	...	0

Figure 1: Table Displaying a Glimpse of the Dataset

This was originally a Kaggle competition and the datasets can be found here. The metric used to determine the success of the models was based on the root-mean-square error.

Data Preparation

Data Preprocessing

Here is an overview of the steps completed in the exploratory data analysis phase:

- Converted Date column to POSIXlt (Date)
- Converted categorical variables to factors
- Created several graphs and tables
    - Bar plots
    - Density plot
    - Revenue histograms
    - Correlation plot
    - Principal component analysis table and visualizations
- Conducted simple feature engineering
    - Log transformed the target variable, Revenue
- Checked model assumptions
    - Linearity, independence, normality, and equal variance
- Removed outliers

Feature Engineering

While exploring the data, it became apparent that revenue had a right-skewed distribution. Thus, a Box-Cox test was done to see if and what transformation would improve the distribution. The results can be seen in Figure 2. Lambda (λ) was -0.14, a log transformation based on the Box-cox criteria. Next, given the results of the Box-cox test, the log transformation was done on Revenue. This made the target variable more normal and, consequently, improved model assumptions. The results of the transformation can be seen in Figure 3.

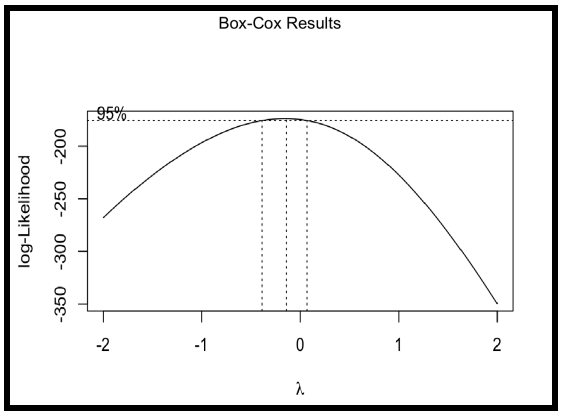

Figure 2: Box-Cox Results

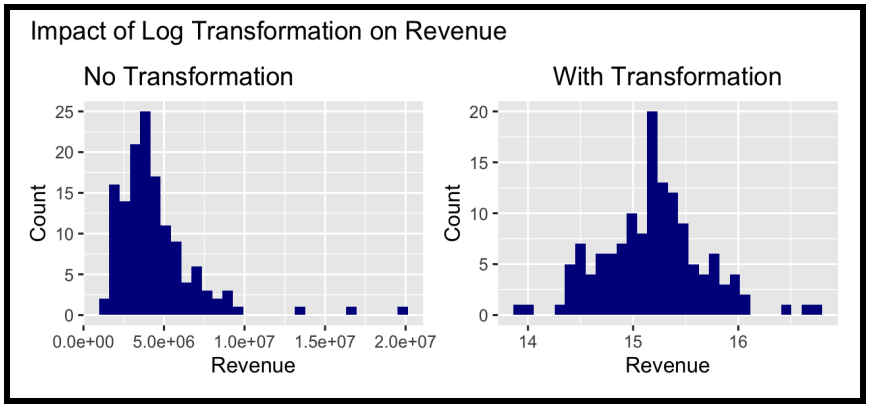

Figure 3: Revenue Histograms

Tidy Model Recipes

The approach used the tidymodels library in R to create various workflows for the models. Creating and processing design matrices for modeling is done by writing recipes. It quickly became apparent that no single recipe worked for all models. However, there were several preprocessing similarities in the recipes that could be used consistently: normalization, dummy variable implementation, removal of zero variance predictors, handling of unseen values, and handling of unknown levels. Robustness checks were completed on each model recipe to make sure additional steps were having a positive impact on the RMSE.

Model Performance

There were 3 main models created in hopes of addressing TFI's concerns: penalized linear regression, light GBM-boosted trees, and random forest. A simple model description and metric results can be found below. Please keep in mind that a lower RMSE is considered better.

Model Performance Comparison

Model Type	Model Description	RMSE
Penalized Linear Regression	Baseline model with auto hyperparameter tuning for penalty and mixture	1837562.77602
LightGBM Boosted Trees	Includes auto hyperparameter tuning for tree_depth, trees, and learn_rate	1770795.43113
Random Forest	Includes auto hyperparameter tuning for min_n and trees	1726644.14318

Figure 4: Model Comparison Table

The penalized linear regression model was used as a baseline because of its interpretability and tendency to avoid overfitting. The auto hyperparameter tuning selected a penalty of 0.026 and a mixture of 0. This is considered ridge regression, it shrinks coefficients toward zero, and helps reduce multicollinearity. A downside of this model is its limited ability in capturing nonlinear relationships. Next, boosted trees are typically high performers in Kaggle competitions because of sequential computation. In this case, the auto hyperparameter tuning selected 2000 trees, a tree depth of 4, and a learn rate of 0.000562 as the best tune. Boosted trees can be sensitive to noise. However, the model that was able to lower the RMSE the most was a random forest. The best tune consisted of 1000 trees and a minimum of 2 data points in a node.

Why Random Forest was Best Performing

There are many reasons why the random forest model may have performed the best. They are listed below:

1. Random forests have a parallel computation approach. Thus, the model aggregates the results of many trees to get an overall prediction. This often helps find a balance between bias and variance.
2. They assume the relationship between the target variable and each feature is nonlinear.
3. Random forests do not require much feature engineering because they naturally can account for interactions.

Conclusion

As previously discussed, Tab Food Investments is interested in predicting the annual revenue of restaurants given certain metrics to inform their decisions on restaurant locations. To discover this information, historical data was used to create a predictive analytics model to forecast future revenue. The profit associated with implementing this model is roughly $1,519,517. This model would allow the company to better allocate funds to restaurant locations with more potential success.

The top 10 significant features from the random forest model can be seen in Figure 5. Interestingly, the 3 most important predictors are the date a restaurant is open, P28, and the city of Istanbul. Date as the most significant predictor logically follows because as a restaurant is more established, it will have increased revenue. Next, all variables starting with P are obfuscated data ranging from demographic, real estate, and commercial data. Without more insight, it is difficult to know which category P28 falls into. Lastly, the city of Istanbul seems to be an important indicator, yet it is important to note that there were 34,087 instances of Istanbul in the test data. To put that into perspective, there were 41 other cities with less than 100 instances and 15 with less than 9,000 instances. This discrepancy may be attributed to the limited data available

for the other cities. While the top three important features may be useful in a predictive model, they may not hold enough practical significance for TFI.

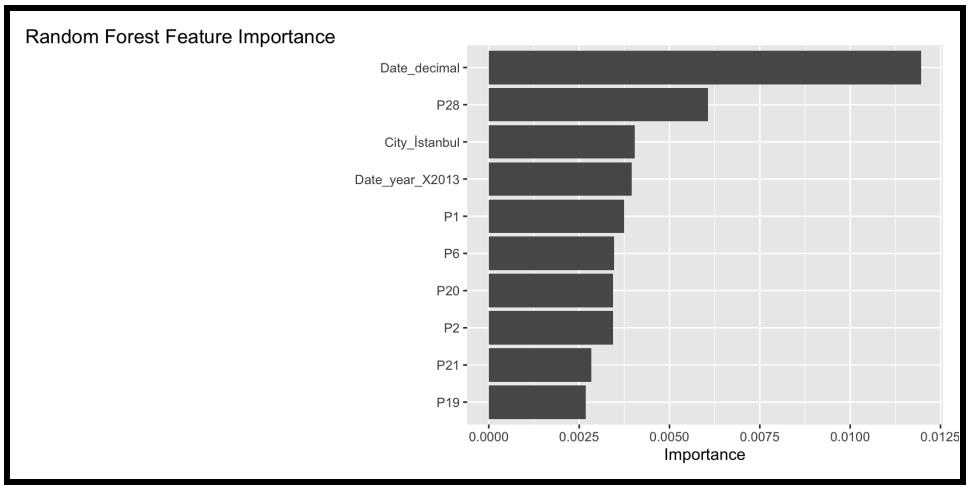

Figure 5: Random Forest Feature Importance for the Test Data

Actionable Insights

1. Implement the random forest model to predict high-performing restaurant locations.
    a. Continue with the expertise of the TFI development teams, but ensure they use this as a main determining factor.
2. Meet with the development team or other stakeholders to determine the practical significance of certain predictors.
    a. As mentioned, all P predictors are obfuscated data. At a minimum, it is essential to share the random forest feature importance with individuals who know what those variables represent. Those individuals can help determine their practical significance. If actionable, TFI could develop more strategies based on those insights.
3. Address data imbalance across cities
    a. TFI should consider adding supplementary data to help minimize the city imbalance. Adding more instances of other cities could also reveal potentially untapped markets.
        i. If after adding more historical data, Istanbul continues to be a top feature, then TFI should consider investigating why Istanbul is such a successful location. A further investigation in this area could allow TFI to identify certain city traits, and find similar cities to open restaurants in. Two approaches to do so could be a cluster analysis or causal inference.
4. Create a target revenue cutoff for restaurant locations
    a. At the moment, this model is purely predicting the annual restaurant sales based on objective location measurements. It may be beneficial to add a cutoff to determine high-performing vs. underperforming locations. This information was never given in the competition.

Limitations

Unfortunately, there are limitations to my conclusions, specifically in the expected profit from the random forest model. First off, TFI never gave criteria on what was considered a good annual revenue vs. poor. When profit was calculated, it was actually a sum of all the predictions for each restaurant. If TFI shared their cutoff, it would be possible to identify the restaurant locations that were considered successful and only calculate those for profit. If they were unable to share this

number directly, it could also be determined from if a restaurant closed and on what date. That is because TFI shared that a poor location typically closes within eighteen months. Other than that, conclusions may not hold with more recent historical data for a variety of reasons. The current training data only ranged from 1996-2014, the latter end which was nearly ten years ago. Figure 6 shows bar plots of several categorical variables and their imbalance. The hope is that more recent data, from 2015 on, would have better class balances. The conclusions above may also not hold if TFI expanded into other countries. As can be seen below, the training data only contains cities in Turkey. While the conclusions may not hold perfectly, the random forest model should be decently generalizable and adaptable to new historical data.

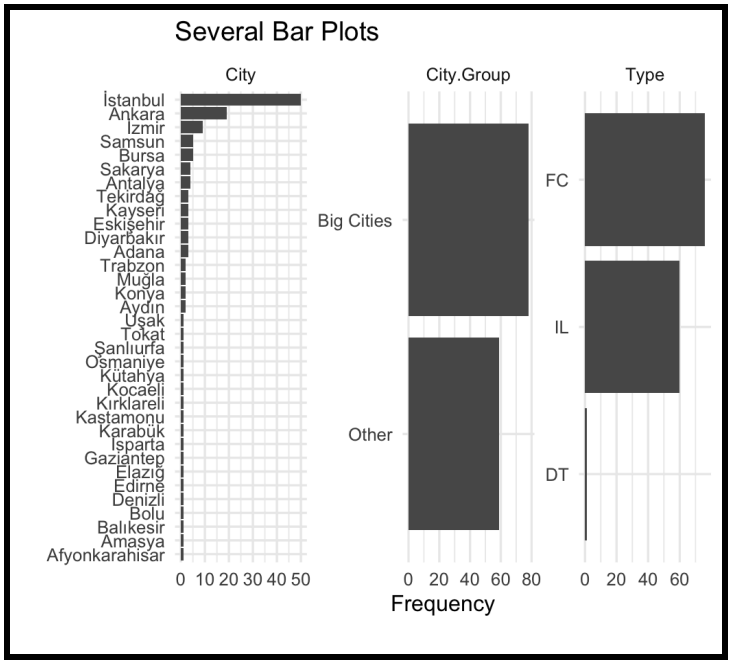

Figure 6: Bar Plots for Several Categorical Variables

Future Directions

Given the limitations mentioned, a great step in the future would be to validate the model against more recent data. It would also be nice to expand the findings through a larger train dataset, if available. This is essential to ensuring model robustness. Next, it may be helpful to get access to what each P variable represents. This information was likely excluded for proprietary reasons, however, it would help interpret the business significance. Specifically, it would likely impact the interpretation of the random forest feature importance. Especially since the top ten predictors mainly consisted of P variables.