Berkeley AI/ML Capstone
05/2025

Restaurant Success and Revenue Prediction Model Analysis

Marcus Lui

What is a measurement of success?

Google Rating?

Annual Revenue?

Average Meal Price?

Social Media Followers?

Years in Business?

Seating Capacity?

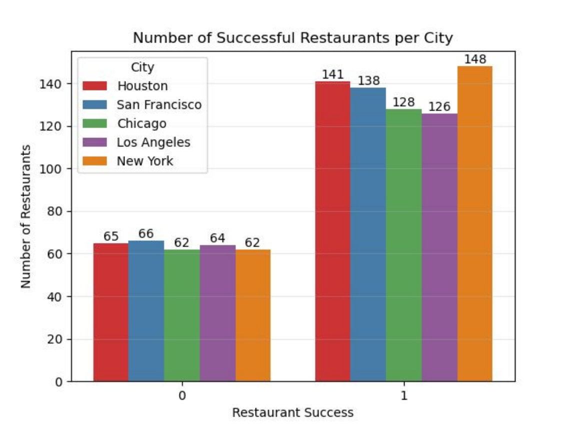
The goal of this project is to identify the best models for restaurant success classification and for revenue prediction

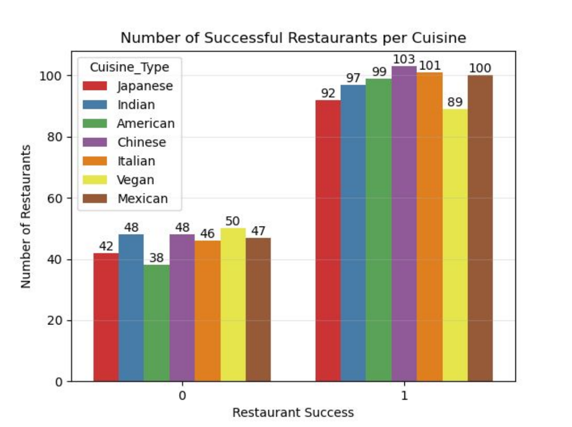
… and to provide information that may help restaurants improve their recipe for success

The *Data*

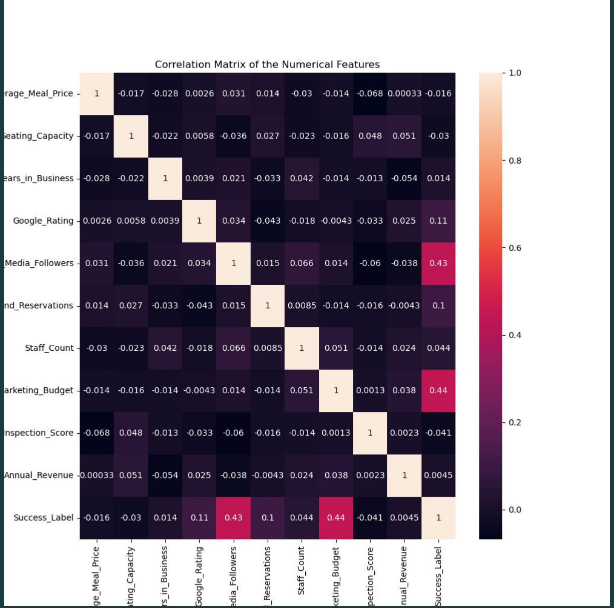
- 4 Categorical Features
  - City
  - Cuisine Type
  - Delivery Service
  - Success Label
- 9 Numerical Features
  - Average Meal Price
  - Seating Capacity
  - Years in Business
  - Google Rating
  - Social Media Followers
  - Weekend Reservations
  - Staff Count
  - Marketing Budget
  - Health Inspection Score
  - Annual Revenue

The First Look



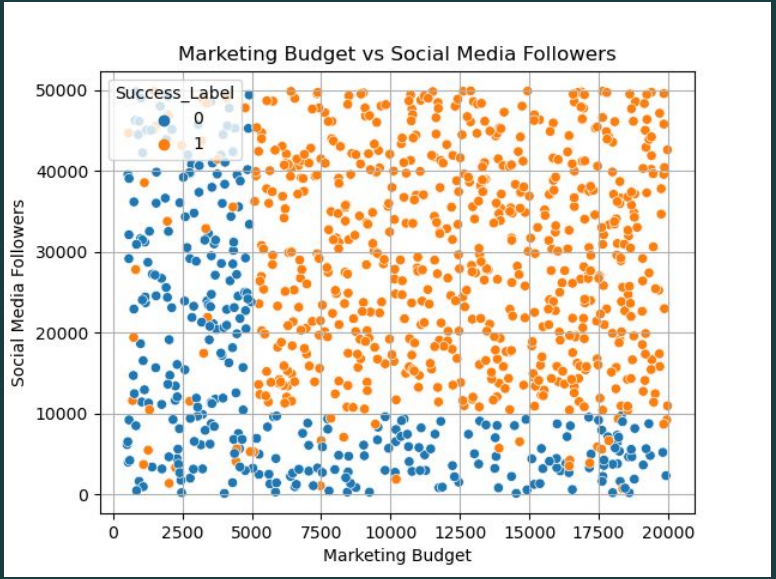


A Deeper Look…



Correlation Matrix of the Numerical Features

	ge_Meal_Price	ating_Capacity	rs_in_Business	Google_Rating	edia_Followers	_Reservations	Staff_Count	keting_Budget	pection_Score	nual_Revenue	Success_Label
rage_Meal_Price	1	-0.017	-0.028	0.0026	0.031	0.014	-0.03	-0.014	-0.068	0.00033	-0.016
Seating_Capacity	-0.017	1	-0.022	0.0058	-0.036	0.027	-0.023	-0.016	0.048	0.051	-0.03
ears_in_Business	-0.028	-0.022	1	0.0039	0.021	-0.033	0.042	-0.014	-0.013	-0.054	0.014
Google_Rating	0.0026	0.0058	0.0039	1	0.034	-0.043	-0.018	-0.0043	-0.033	0.025	0.11
Media_Followers	0.031	-0.036	0.021	0.034	1	0.015	0.066	0.014	-0.06	-0.038	0.43
nd_Reservations	0.014	0.027	-0.033	-0.043	0.015	1	0.0085	-0.014	-0.016	-0.0043	0.1
Staff_Count	-0.03	-0.023	0.042	-0.018	0.066	0.0085	1	0.051	-0.014	0.024	0.044
arketing_Budget	-0.014	-0.016	-0.014	-0.0043	0.014	-0.014	0.051	1	0.0013	0.038	0.44
nspection_Score	-0.068	0.048	-0.013	-0.033	-0.06	-0.016	-0.014	0.0013	1	0.0023	-0.041
Annual_Revenue	0.00033	0.051	-0.054	0.025	-0.038	-0.0043	0.024	0.038	0.0023	1	0.0045
Success_Label	-0.016	-0.03	0.014	0.11	0.43	0.1	0.044	0.44	-0.041	0.0045	1



Marketing Budget vs Social Media Followers

The correlation heatmap (left) shows the features that shows some direct effect on restaurant success and is then confirmed via the scatterplot (above).

Model Types

Classification Models

- Dummy
- Logistic Regression
- K-Nearest Neighbors
- Decision Tree
- Random Forest

Regression Models

- SVM
- Linear Regression
- Lasso
- Random Forest

Classification Models

Logistic Regression (LogReg)

A statistical model used for binary classification

K-Nearest Neighbors (KNN)

A simple, non-parametric algorithm used for classification

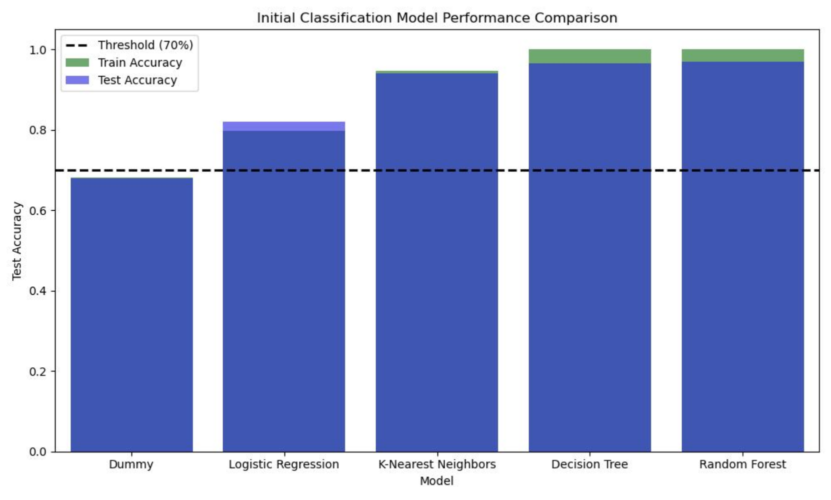
Decision Tree (DT)

An algorithm that recursively splits data into branches based on the features that provide the best separation until all nodes belong to a class

Random Forest (RF)

An ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and prevent overfitting

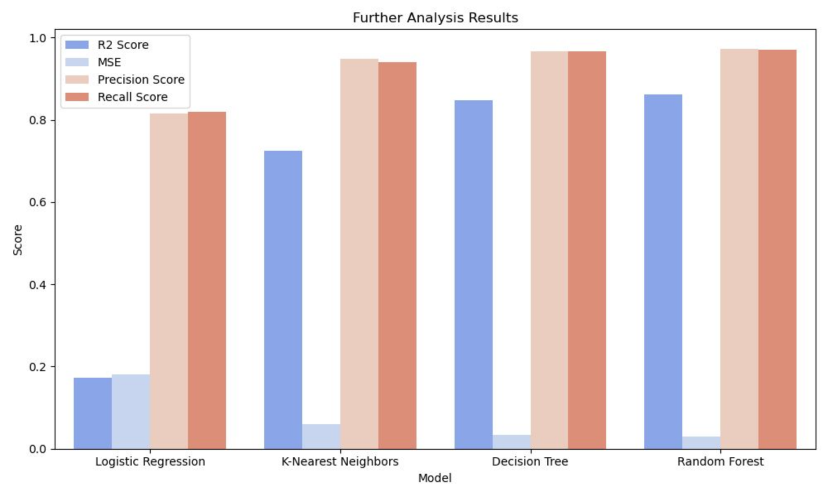
Classification Model Analysis



LogReg and KNN models performed better than expected, recording a 82% and 94% test accuracy, respectively.

GridSearchCV was then applied to determine the best hyperparameters per each model. Results from the enhanced models were similar to the initial model performances.

Decision Tree and Random Forest models both recorded a 100% training accuracy, while both testing accuracies regressed by ~3%



KNN, DT, RF all had a precision score and recall score above 90%. These two models also had a low MSE and a high R2 Score.

For this dataset and this task, DT and RF seem like good models to use for predicting success.

Precision Score

It answers the question:

“Of all the positive predictions made by the model, how many were actually correct?

MSE

It measures the average squared difference between the actual (true) values and the predicted values. Low MSE means higher prediction accuracy.

R2 Score

It is a measurement of how well the models is at explaining variability

Regression Models

Linear Regression (LinReg)

A fundamental machine learning algorithm used to model the relationship between a dependent variable and one or more independent variables

Lasso

A Linear Regression technique that should help with feature selection and thus reduce overfitting

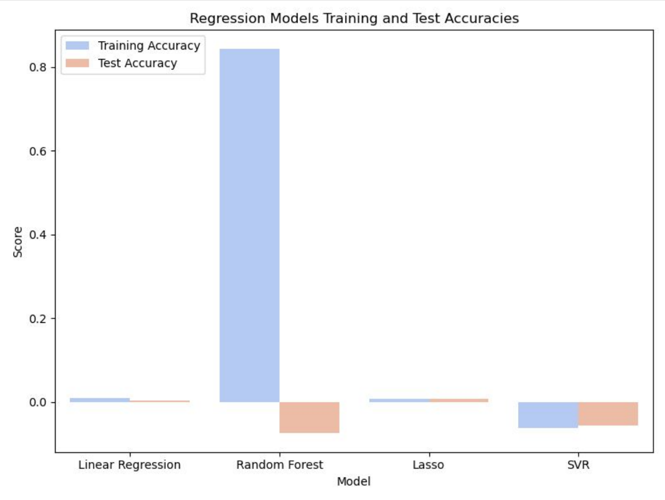
Support Vector Regression (SVR)

Predicts continuous values by finding a function that fits the data within a certain margin of tolerance, ignoring outliers

Random Forest (RF)

An ensemble machine learning method that uses multiple decision trees to predict a continuous output

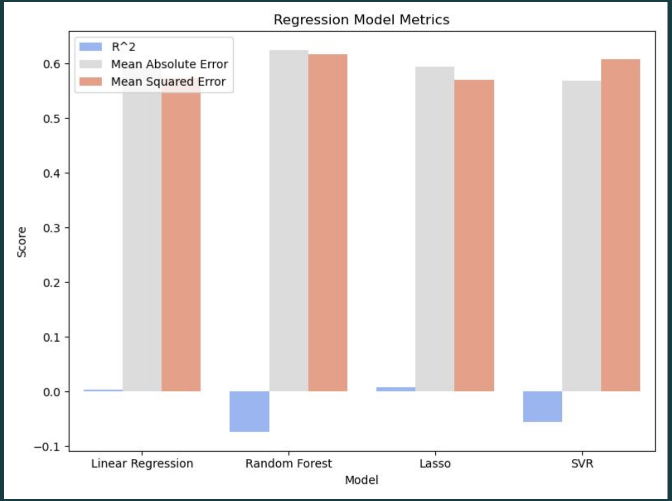
Predictive Model Analysis



Unlike the results of the Classification models, the Linear Regression (LinReg), Random Forest (RFR), Lasso, and SVR models all had a low testing accuracy.

GridSearchCV was then applied to determine the best hyperparameters per each model. Results from the enhanced models were no different to the initial regression model performances.

Such drastic results suggests the data is too scattered, and an unbiased model that ensures high accuracy and high precision cannot be created



LinReg, RFR, Lasso, and SVR all seemed to have a high MAE and MSE score while also having a low R2 score.

For this dataset and this task, there doesn’t seem to be any models that have a high accuracy to use for predicting annual revenue.

MAE

It is the average of the absolute differences between the predicted and actual values. Lower MAE means higher prediction accuracy.

MSE

It measures the average squared difference between the actual (true) values and the predicted values. Low MSE means higher prediction accuracy.

R2 Score

It is a measurement of how well the models is at explaining variability

Model Results Summary

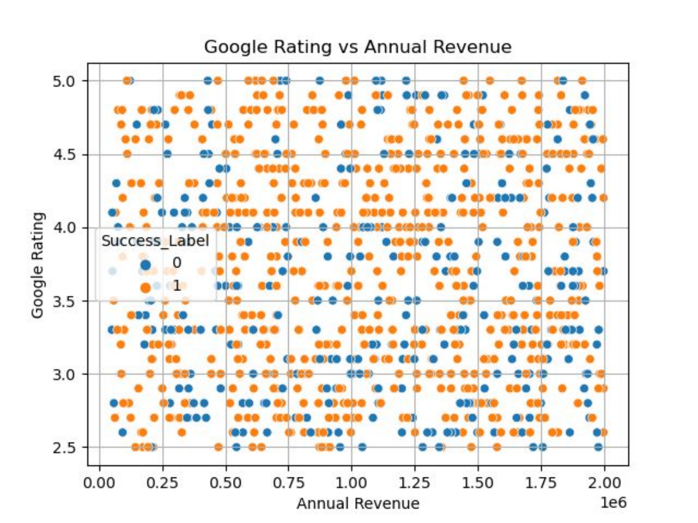
With DT and RF, you can accurately classify a restaurant as successful or not successful with over 95% accuracy.

The results of the DT and RF models have little variance and low error and thus is proven to be both precise and accurate

RFR had the best training accuracy, but recorded a surprisingly low test accuracy, which suggests that the models may be over biased.

All regression models tested performed poorly on this data, implying that the features available do not capture the complexity of what drives annual revenue

Key Takeaways



Use Social Media and Boost Marketing

Restaurants that prioritized a higher marketing budget and a greater online presence were recognized as successful.

Revenue Prediction Requires More Quality Data

The dataset used had both categorical data and numerical data, but a limited amount. Good data is better than noise

New Features and Metrics

Features that could improve the dataset include customer loyalty data, seasonal trends, or competition density

Thank you

https://github.com/MarcusLui9/Restaurant_Success_and_Revenue_Prediction
https://www.linkedin.com/in/marcus-chi-kin-lui/