Predicting Restaurant Revenue Based on Performance Metrics

Project Outline

Introduction

- Give what, why and how:
    - What: predict monthly revenue
    - Why: gives necessary data to franchise owners / food service industries to make informed business decisions
    - How: use python Jupyter notebook to produce regression models that minimize RMSE for expected revenue
- Explain in more detail examples of restaurant franchises (such as Yum! Foods or RBI) that could use this information, and what types of decisions this would help inform

Dataset

- Cite the Kaggle synthetic dataset (https://www.kaggle.com/datasets/mrsimple07/restaurants-revenue-prediction) inspired by the Kaggle competition (https://www.kaggle.com/competitions/restaurant-revenue-prediction)
- Introduce each of the variables, what they represent, and what the units of measure are
    - Units are not explicitly given, so from just making an educated guess I'd say
        - # of customers = avg. daily customers
        - Avg. menu price = avg. # of dollars for entrée items
        - Avg. spending per customer = avg. # of dollars spent on non-entrée items
        - Monthly revenue & marketing spending = measured in thousands of dollars
    - Let's say that the restaurant also makes money from deliveries, fees, and merchandise that is not included in the dataset (which is why you can't just do [customers + spending]*30 = revenue), or that maybe this data only includes a certain portion of customers and not every customer

Model Selection

- Models:
    - Linear regression – baseline

    - L1 and L2 regularization (lasso and ridge)
    - Random Forest Regression
- Explain what each of these models do, why they would be appropriate for this dataset, and how they will be implemented

Evaluation Method

- Split data into training and test sets, using one-hot encoding for categorical data
- K-fold cross-validation + hyperparameter tuning
- R^2 and RMSE will be the performance metric of each model
- Explain why we are using these^

Learning Objectives

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

Project Risks

- Synthetic data: since the data is machine-generated and not necessarily based on a real-world company, we may not be able to use to it to make meaningful generalizations
- Overfitting: since the dataset is not particularly large, we will need to be careful of overfitting, especially for random forest classification
- Multicolinearity: It may be the case that the number of customers, average spending, reviews and other factors have a high degree of correlation, and this should be identified before further analysis

Ethical Implications

- Since this is machine-generated data, we should not ethically use it to make broad statements about typical real-world restaurants or franchises
- We should not monetize this research or use it for any financial benefit
- The assumptions we have made about the dataset and what it represents may be inaccurate, and we should be careful so as not to misrepresent what we have stated as absolute fact

Contingency Planning

- If this data proves to be underfit for effective analysis, we may need to seek out additional data, or another data source altogether
- If the data is too overfit, or the models are very inaccurate, we should adjust hyperparameters or seek alternative models
- If there is too much variability, there may be confounding variables that should be taken care of, as well as outlier elimination

Sources

- Use APA format and cite the textbooks, online articles / journals, and Kaggle