Project: To predict the annual restaurant sales of 100k regional locations

Abstract

Food/Restaurant Industry is one of the evergreen Business fields and it has showed more of its potential in the recent days. Nowadays, Customers has got a wide range of options for Food based on the Cuisine type, Food quality, Taste and Varieties and they always looks for something new and better in every aspect. So, this industry can expect a tremendous growth in the future also. Large Capital investment, Effort and Time are required for setting up and run a restaurant smoothly. Therefore, it is crucial to consider a lot of factors while launching a restaurant anywhere. Our goal is to make use of the power of Data through various Machine Learning, Deep Learning techniques to increase the effectiveness of investments in new restaurant sites and to gain an upper hand in the competitive business.

Business Problem

Introduction

Tab Food Investments (TFI) is the company behind some of the world's most well-known brands such as Burger King, Sbarro, Popeyes, Usta Donerci and Arby's, with over 1200 quick service restaurants across the globe. They employ over 20,000 people in Europe and Asia and they are the Leading Quick Service Restaurant (QSR) operator in Turkey and China.

Problem Statement

Right now, new restaurants are launched in a particular location is commonly a subjective process based on the personal judgement and experience of development teams. This subjective data is hard to meaningfully extrapolate across different geographies and cultures.

It takes large investments of time and capital to set up new restaurants and running. The site closes within 18 months by incurring operating losses, when the wrong location for a restaurant brand is chosen.

We are expected to find a mathematical model to increase the effectiveness of investments in new restaurant sites which would allow TFI to invest more in other important business areas, like innovation, sustainability, and training for new employees. We would try to predict the Annual Restaurant sales of 100,000 regional locations based on the given Commercial, Demographic and Real Estate data.

Dataset

Source of the Dataset

The TFI Dataset for predicting Restaurant Revenue prediction is available on Kaggle. Both the train and test dataset are provided here.

The dataset can be downloaded from the following link:

https://www.kaggle.com/competitions/restaurant-revenue-prediction/data

Overview

The TFI Dataset consists the Train and Test Dataset files. There are 137 entries of Restaurants in the Train dataset which will used for training and fitting the model. The Test dataset consists of 100,000 samples of data which will be used for the evaluation of the model. We can observe that the train dataset provided is very small compared to the test dataset.

The Test Dataset has similar features of the Train Dataset except that the Revenue column can be only found in the train data. The features of the dataset are described below:

- ID – Restaurant Id
- Open Date- Date of Lauch of the Restaurant
- City- The city name where the Restaurant situates
- City Group- The category of the city which the restaurant belongs.
- Type- The type of the Restaurant with categories as Food court, Drive-Thru, Mobile, Inline.
- P variables (P1-P37)- These variables represent the obfuscate data of three categories:
    1. Demographic data - Population, Age, Gender etc.
    2. Real Estate data – Car parking availability, Front Façade etc.
    3. Commercial data - Points of Interest such as Colleges, Market etc.
- Revenue- Annual Revenue of a Restaurant in a given year.

So, there are 43 features in the Train dataset including the Restaurant id. Since the revenue column is excluded in the test data it has 42 attributes in total.

Challenges

The difference in the size between the Train and Test dataset is huge. The fact that the train dataset has only a small number of data points would be a challenging part to deal with, from the Data modelling perspective. There is big probability that the models can overfit easily, especially the complex ones, which can negatively impact in the prediction of the revenue. To prevent this, we would be needed to use regularization techniques.

Also, the dataset contains some missing values, categorical features. These data have to be preprocessed before going ahead with the Modelling. We would be using the different libraries like pandas, scikit-learn for Data Preprocessing.

Key Performance Indicator

In this problem, we will be trying to predict the Annual revenue of the restaurants. Since the output we are predicting is a real/continuous value, this is a Regression type problem. For Regression models typically we employ some performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE) etc.

We can evaluate the performance of the developed model by comparing the Predicted revenue to the Actual revenue given in the dataset. Here, we can implement RMSE as the performance evaluation metric.

Root Mean Squared Error:

RMSE is one of the most popular evaluation metrics for Regression problems. It is the square root of the average squared errors/residuals. The mathematical formula of RMSE is given below:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

where,

Y_i = Actual value

Y_i hat= Predicted value

n=no. of datapoints

Lower the RMSE value implies the model performs better.

The key advantage of using RMSE as the metric is that the unit of the error, we obtain will be same as that of the output variable that we predict. In our case, if the revenue is expressed in the unit *Dollars* then the error or residual value obtained will be also in *Dollars.* This would make the interpretation easier.

The Disadvantage of RMSE is that it is not much robust to the outliers. So, it is critical to process or handle the outliers when we are employing this metric for evaluation.

Alternative metric that can be used:

Mean Squared Error (MSE) is one of the alternative metrics that can be used for the Regression problems. It is simply the average of the squared residuals.

$$MSE = \frac{1}{n} \Sigma \left(\underbrace{y - \widehat{y}}_{\text{The square of the difference between actual and predicted}} \right)^2$$

The prime reason that we are not using the metric is because of it less interpretability in results, since the value that we get after calculating MSE is the squared unit of the actual output variables.

Applications on Real world problems:

The RMSE performance metric can be used for all supervised learning applications where the predicted outputs are real/continuous values.

Real World Challenges and Constraints

Overfitting

The train dataset has only a small number of datapoints. So, there is a high chance of overfitting the datapoints, especially when we run a complex model. The test dataset which is of 100,000 datapoints is about 729 times the size of the train dataset. Therefore, an overfitting model could lead to worst performance of the model on the test data.

In order to overcome this problem, the model should be well selected and tuned. Since large investments are to be made for setting up a new restaurant, it is a big challenge to develop a well optimized model to provide more accurate predictions to avoid losses and generate good profits.

Obfuscated variables

Feature interpretation is a very important factor in decision making. Since most of the variables are obfuscated, it is difficult to interpret the meaning of this features. Moreover, these variables are generated from the demographic, commercial and real estate, and logically thinking this variables/factors would be super important in the case of our problem. So, the abstraction of these feature is a challenge to de dealt with.

References

- https://www.kaggle.com/competitions/restaurant-revenue-prediction/data
- https://hpriya206.medium.com/sales-prediction-using-python-for-machine-learning-6a76e4d63e71
- https://www.academia.edu/download/49081314/J0604091094.pdf
- https://www.irjet.net/archives/V8/i12/IRJET-V8I1210.pdf
- https://saiharishcherukuri.medium.com/restaurant-revenue-prediction-ddca2ed65da7
- https://medium.com/hamoye-blogs/restaurant-revenue-prediction-f266a974c6b5
- https://towardsdatascience.com/restaurant-revenue-prediction-467f0990403e
- https://www.youtube.com/watch?v=YiXeSJtEgNw