THREE DIFFERENT MODELS FOR RESTAURANT REVENUE PREDICTION

Jiedong Tong, z5176253

August 8, 2020

Introduction

TFI owns more than 1,200 fast food restaurants worldwide and is the company behind some of the most famous brands in the world: Burger King, Sbarro, Popeyes, Usta Donerci and Arby's. They have more than 20,000 employees in Europe and Asia, and invest heavily in the development of new restaurants every day.

Currently, based on the personal judgment and experience of the development team, deciding when and where to open a new restaurant is largely a subjective process. These subjective data are difficult to accurately infer various regions and cultures.

The opening and operation of a new restaurant requires a lot of time and capital investment. If the wrong location is selected for the restaurant brand, the website will be shut down within 18 months and incur operating losses.

After preprocessing the data, I chose three different mathematical models for machine learning and predict the revenue in different situations.

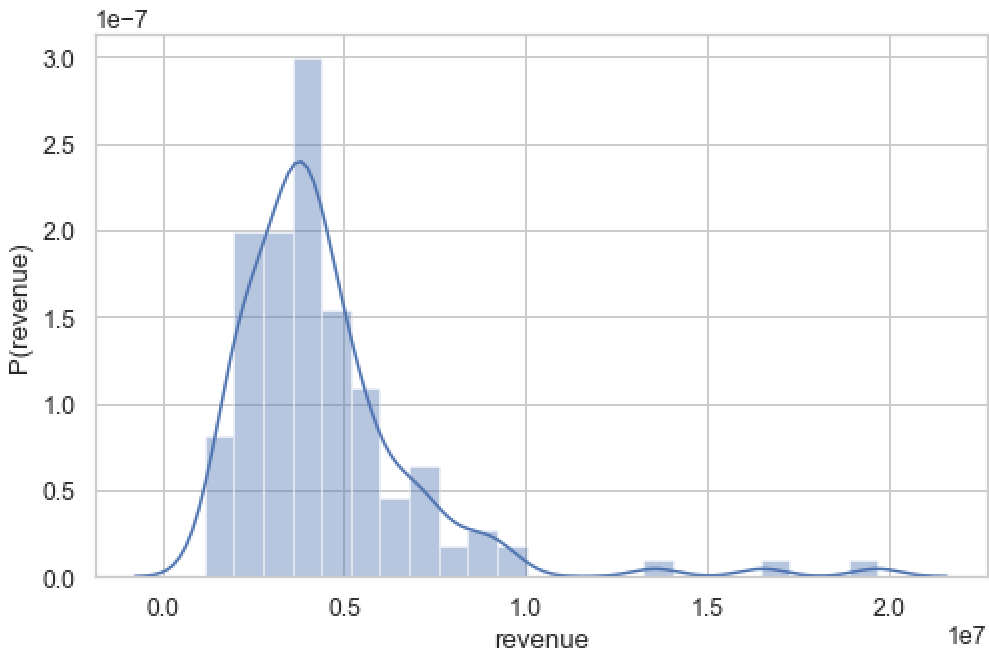
The datasets for Restaurant Revenue can be download here :
https://www.kaggle.com/c/restaurant-revenue-prediction/overview/description

All codes, sample output csv files and all used datasets can be download at the following location : https://github.com/zzwzs/comp9417project/tree/master/project

Overview of datasets and preprocessing

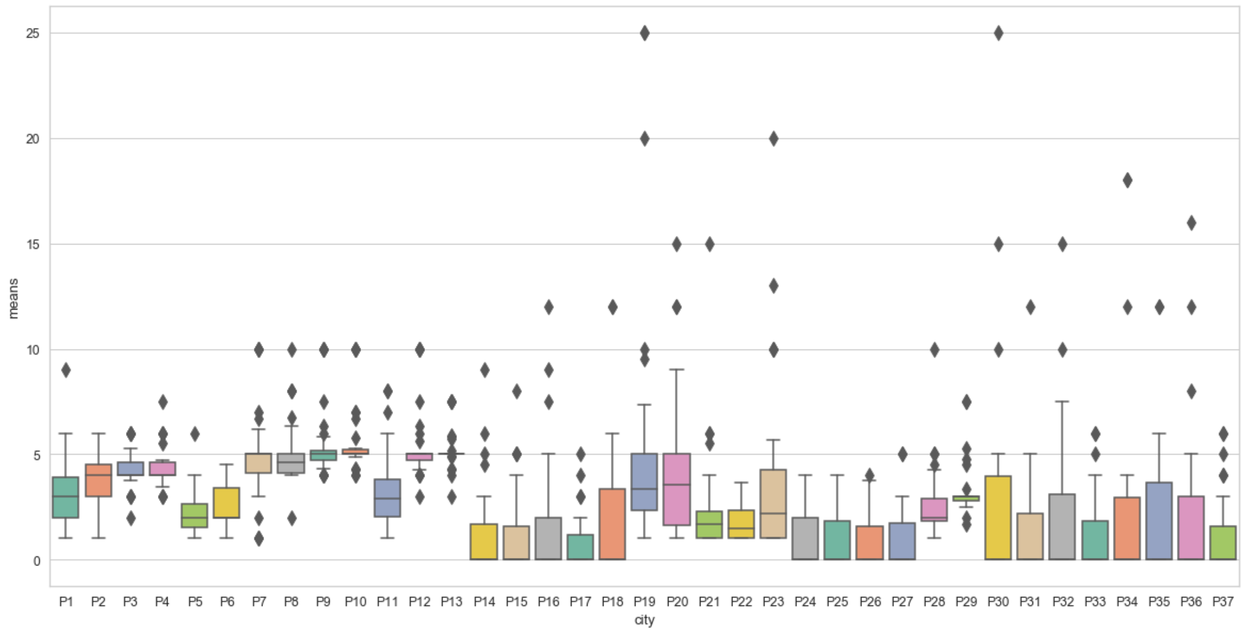
For the entire dataset, there are generally out layers, and these out layers are generally generated by some special circumstances or accidentally. If these out layers are included, the prediction accuracy of mathematical models will be significantly reduced.

I drew the train dataset into a column chart as below:



Obviously, we can find that if revenue greater than 1*10^7 then it may be the out layer. Because most of the revenue is less than 1*10^7, only a few of revenue not. Therefore, those out layers should be deleted.

The degree of influence of different features on prediction may also vary greatly, but if the train set is not very substantial, it is difficult to determine the weight of different features. A better way is to extract the features that have a greater impact on the prediction and discard the features that have less impact. I plot all the ‘p’ features and the average of the ‘p-value’ generated as below.



According to this graph, I find that 'P1', 'P2', 'P8', 'P11', 'P19', 'P20', 'P23','P24', 'P30' have a greater impact on the prediction.

Extra features

1. Cluster.
   clustering algorithm comes in two variants: a class, that implements the fit method to learn the clusters on train data, and a function, that, given train data, returns an array of integer labels corresponding to the different clusters. I tried many different K-Means value (the number of clusters) and find that 25 fits here.
2. City_G (I separated City Group by different value)
   Looks like City_G_BigCities and City_G_Other (if the value here is 0 that means this city is not belong to this class)
3. type (I separated Type)
   type_DT, type_FC, type_IL, type_MB (if the value here is 0 that means type is not same as this type)
4. Open date

Year, Month, Day

Three different models

1. Decision Tree

   A decision tree is a structure similar to a flowchart, where each internal node represents a "test" on an attribute (for example, whether a coin flip is heads or tails), each branch represents the result of the test, and each leaf node represents a test result . Class label (decision made after calculating all attributes). The path from the root to the leaf represents the classification rule.
2. Neural Network

   Artificial neural networks are based on a collection of connected units or nodes called artificial neurons, which can loosely model neurons in a biological brain. Each connection is like a synapse in a biological brain, which can transmit signals to other neurons. The artificial neuron that receives the signal then processes it and can send a signal to the neuron connected to it.
3. Support Vector Machine

   In machine learning, a support vector machine (SVM, also known as support vector network [1]) is a supervised learning model with related learning algorithms that analyze data used for classification and regression analysis

References

Support vector machine, Wiki, https://en.wikipedia.org/wiki/Support_vector_machine

Artificial neural network, Wiki, https://en.wikipedia.org/wiki/Artificial_neural_network

Decision tree, Wiki, https://en.wikipedia.org/wiki/Decision_tree

Clustering, Scikit-learn, https://scikit-learn.org/stable/modules/clustering.html

Dummies, Pandas, https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html

2015, Restaurant Revenue Prediction - Predict annual restaurant sales based on objective measurements , https://www.kaggle.com/c/restaurant-revenue-prediction/overview