Restaurant Revenue Prediction Problem

1. Description of the data

1.1 Data Source and Purpose

The data comes from the TFI Restaurant Revenue Prediction Challenge, which is a competition hosted on Kaggle. The purpose of this challenge is to predict the annual sales of restaurants based on given objective measurements.

1.2 Variables Description

The whole data includes a training data set that contains 137 entries and a test data set that had 100,000 entries. The data contains 4 categorical variables and 37 ordinal anonymous variables. The categorical variables include city name, city group (big or small), type of the restaurant (FC: Food Court, IL: Inline, DT: Drive Thru, MB: Mobile) and opening date of the restaurant. The ordinal anonymous variables are P1, P2, . . . , P37. These variables capture various parameters relating to demographics, real estate, and other commercial evaluations. The revenue column in the training dataset indicates a transformed value of the actual revenue of restaurants, and is the target of the predictive analysis.

1.3 Data Preprocess

1.3.1 Transform Open Date

We compute a variable named ‘time’ from the Open.Date column provided by subtracting Open.Date from system date.This is the days that a restaurant has been open. Since the time variable has a skewed distribution, a log scaling on it gives a much better distribution. A log also deals with outlier problem.

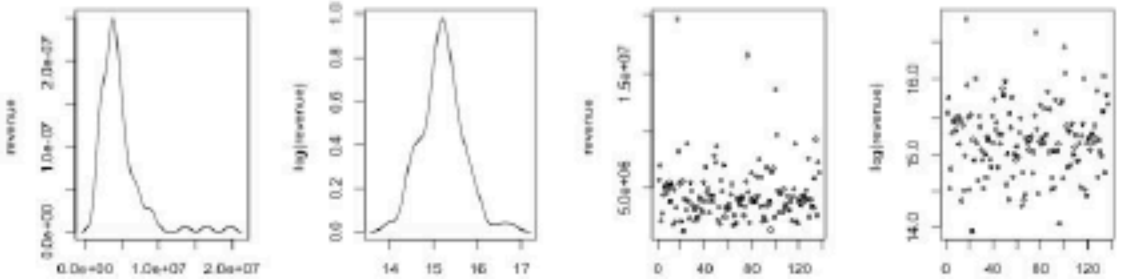

Figure 1: Scatter plots and distribution plots of time and log(time)