Restaurant revenue prediction

Team 38

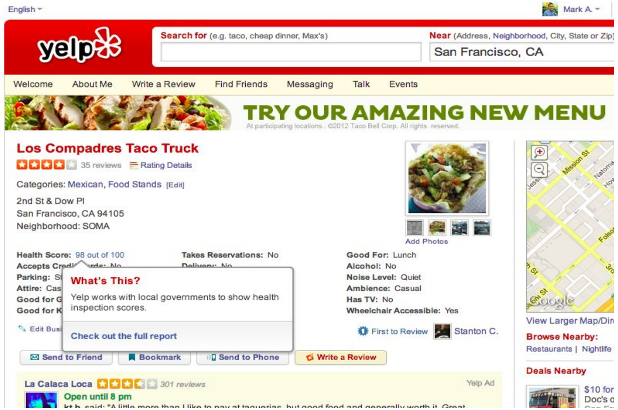
Anjali Shenoy Rehas Sachdeva Saumya Rawat

Background

Enterprises today, highly motivated in the art of finding **anomalies, patterns and correlations** within data-sets.

Want to improve their **online reviews to attract clientele** or seek to establish a new business that is mindful of what **drives good reviews**, particularly true for **restaurants and food establishments.**

TFI TAB FOOD INVESTMENTS



Search for (e.g. taco, cheap dinner, Max's)

Near (Address, Neighborhood, City, State or Zip)

San Francisco, CA

Los Compadres Taco Truck

35 reviews Rating Details

Categories: Mexican, Food Stands [Edit]

2nd St & Dow Pl
San Francisco, CA 94105
Neighborhood: SOMA

Health Score: 98 out of 100
Accepts Cred
Parking: St
Attire: Cas
Good for G
Good for K

Takes Reservations: No
Delivery: No

Good For: Lunch
Alcohol: No
Noise Level: Quiet
Ambience: Casual
Has TV: No
Wheelchair Accessible: Yes

What's This?

Yelp works with local governments to show health inspection scores.

Check out the full report

First to Review Stanton C.

La Calaca Loca 301 reviews

Open until 8 pm

kt h. said: "A little more than I like to pay at taquerias, but good food and generally worth it. Great

$10 for
Doc's o

Introduction

Several studies conducted to look at the **correlation between a restaurant's success and its reviews and ratings.**

TFI is behind the famous brands like Burger King, Sbarro, Popeyes etc, interested in **extrapolating their data across geographies and cultures.**

We will be working with a **TFI data set of about 1 lakh Turkish restaurants.**

Problem Statement

- Supervised learning problem, **objective is to develop a model and a set of preprocessing procedures** to accurately **predict the annual restaurant sales** of 100,000 regional locations using various parameters.

- Given are 43 attributes, 5 trivial ones like restaurant type, city etc and a group of **37 obfuscated P-variables** like population, parking availability, other vendors etc. The annual revenue attribute is to be determined.

- There are other inherent problems like **unaccountability** of attribute ranges, **categorical vs continuous problem and zero problem**.

Solution

- Based on **Random Forest** and **Support Vector Machine**.

- Preprocessing - **analysis of histograms** for number of restaurants in a revenue range to try to find some underlying distribution or conversion to get something like a normal distribution.

- Unaccountability problem, we can use **KNN** or **K means**, depending on which performs better, to match records with an unaccounted value to one with accounted value. **DB Index plot** to solve the continuous vs categorical problem, and **KNN** again for the zero problem. **PCA** to reduce the dimensionality of the data especially the 37 P-variables.

- Apart from all this mentioned in the paper, we would implement a **neural network based model**. The Kaggle competition had a constraint of 137:100000 training to test data ratio. But we have access to both. So we can divide the data more evenly and implement a **K-fold cross validation** as well. We would also compare the results of all the models.

- We can visualize results using **t-SNE technique**. We will also try to interpret the results as in find which parameters are highly determining the revenue and visualize the same.

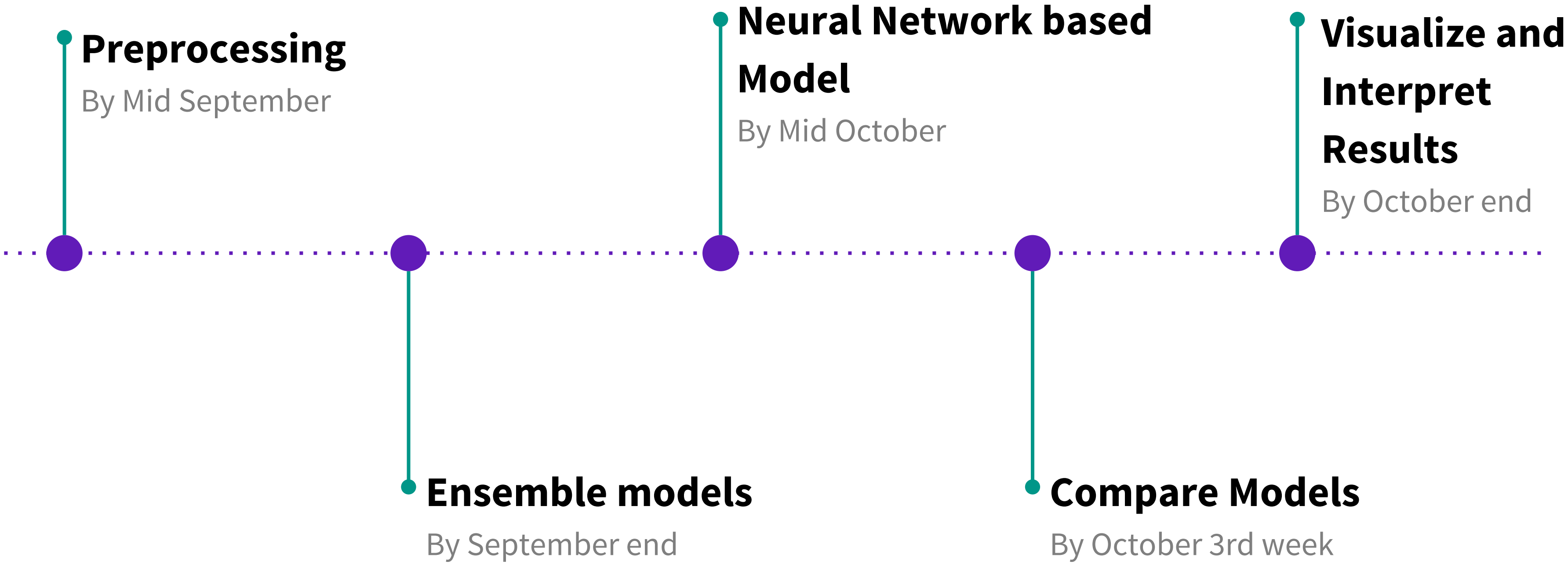
Languages and Toolkits

- R Language
- MATLAB
- SAS
- JMP
- t-SNE

Scope

- The solution is **only applicable to Turkish restaurants** and locations on which the data is based. A different location based data may need a different kind of ensemble for accuracy.
- It is limited to only **annual and not seasonal** revenue analysis.
- Our scope is also **limited by the neural network model** that we build.

Timeline



Preprocessing

By Mid September

Ensemble models

By September end

Neural Network based Model

By Mid October

Compare Models

By October 3rd week

Visualize and Interpret Results

By October end