Statistical Methods in Artificial Intelligence

Team 38: Restuarant Revenue Prediction

Anjali Shenoy (201401114), Rehas Sachdeva (2014010..), Saumya Rawat (201401110)

BACKGROUND & INTRODUCTION:

Enterprises today, highly motivated in the art of finding **anomalies, patterns and correlations** within data-sets. These enterprises seek 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.
Several studies have been 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:

"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."

We classify this problem as a supervised learning problem.

DATASET DESCRIPTION:

★ The dataset represents ~100,000 Turkish restaurants as uploaded on.

★ Kaggle. The size of the training dataset is 137 whereas the size of the validation set is almost 100,000.

★ The 43 various attributes of the dataset are:

- **Id**: Restaurant id.
- **Open Date**: Date that the restaurant opened in the format M/D/Y
- **City**: The city name that the restaurant resides in
- **City Group**: The type of city can be either big cities or other
- **Type**: The type of the restaurant where **FC** - Food Court, **IL** - Inline, **DT** - Drive through and **MB** - Mobile.
- **P-Variables (P1, P2, ... ,P37):** Obfuscated variables within three categories: demographic data, e.g population, age, gender; real estate data e.g car park availability and front facade; commercial data e.g points of interest, other vendors, etc.[1]
- **Revenue**: Annual (transformed) revenue of a restaurant in a given year and is the target to be predicted.

FEATURE EXPLANATION:

- Sales depend on the location and type of city it is in. If the city is a metropolitan city, it will have a larger customer base than a town, and hence revenue generated will be different in both these cities.
- On a similar note, revenues generated will be different for different restaurant types- Drive through will attract more customers in a remote area where as a restaurant will attract more customers if it is placed in the heart of the city.
- The open date attribute doesn't do much as such. But if processed to get number of days a restaurant stays open, year of opening, month of opening, we can get an idea of cyclical or seasonal patterns that affect revenue.
- Apart from these, we have demographic attributes e.g population, age, gender; real estate data e.g car park availability and front facade; and commercial data e.g points of interest, other vendors, etc. These also decide sales to varied extents. They can also be correlated

[1] It is unknown if each variable contains a combination of the three categories or are mutually exclusive.

CHALLENGES:

1. The size of training dataset is 137 samples while that of test dataset is 1,00,000 samples. This is a **large disparity**.
2. **We don't have the ground truth for our test data**. So we cannot use sophisticated performance measures like precision, recall, k-fold cross validation etc. We only know RMSE for the entire test data.
3. Whether to predict revenue or log(revenue) as we see that training data follows normal distribution when taken with log(revenue). But we can't say the same about test dataset.
4. **Parsing** the data types of various attributes.
5. **Unaccountability problem:** the disparity between the features for the training set and test set as the test set contains more information than the training set.
6. **Categorical vs continuous problem:** whether the obfuscated P-Variables should be treated as categorical or continuous.
7. **Zero problem:** For certain P-Variables, a large number of samples contain zero values and are dependent among each other such that if one p-variable has zero on a certain row, the probability that other p-variables take on a zero value is high.

LANGUAGES & TOOLKITS:

- Python
- SKLearn toolkit

FEATURE EXTRACTION & SELECTION:

The training and test data are available as CSV files, containing a total of ~100,000 samples of Turkish Restaurants. The input data is too large to be processed and majority of the data fields are obfuscated variables without giving any prior knowledge of each one.

Our pre-processing is done as follows:

1. We use histograms to see that log(revenue) follows an approximately normal distribution. So choosing target variable as log(revenue) instead of revenue improves performance of base models.
2. Opening date cannot simply be assumed to be a factor so two additional features are created: month that they opened and the year that they opened. These two features can potentially help proxy seasonality differences since restaurant revenues are highly cyclical.
3. Since the restaurant 'Type'- 'MB' is not present in the training but available in the Test, the Test set is modified so each mobile type restaurant is matched with a non-mobile type restaurant through finding the most similar features, as measured by euclidean distance.
4. Similarly for 'City' since the number of cities in the test set is more than the training set, using KNN all 137 cities are clustered on the basis of P Variables that best describe Geographical locations and then these clusters are assigned to the city column of the datasets. To know exactly what each p-variable represents, under the assumption of mutually exclusive categories, a change in the mean over each city should elicit a change in certain p-variables. And using box plot of mean p-variables over each city, P1, P2, P11, P19, P20, P23, and P30 are identified to be approximately a good proxy for geographical location.
5. We also use PCA for the P-Variables because we aren't given exactly what these represent. They can be correlated. So PCA can represent them better.

VALIDATION TECHNIQUES USED:

We use histograms to see that log(revenue) follows an approximately normal distribution so choosing target variables as log(revenue) instead of revenue improves performance of base models.

We also visualize a box plot of mean of P-variables for each city to help us identify which P-variables are majorly geographical.

We also apply Davies-Bouldin index for K-Means clustering on variables: P1, P2, P11, P19, P20, P23, P30 to help us validate the best K to be used as number of clusters.

PERFORMANCE METRICS:

Since we **do not have the ground truth for the validation set**, we mainly tested out performance **by submitting our code on kaggle** to see what rank we got. We also had the RMSE values for the dataset and hence tried to use this as a parameter for minimisation.

ALGORITHMS:

1. KNN to account for types of restaurants that are there in test data but not in training data. The idea is to match those unknown types with known types (from training data) using nearest neighbour method based on rest of the attributes. KNN is also used for zero problem, to get an appropriate value for missing values of various attributes.
2. Kmeans to account for cities of restaurants that are there in test data but not in training data. The idea is to match those unknown cities with known cities (from training data) using clusters of cities.
3. PCA to reduce the dimensionality of the data, especially useful for the p variables, because we are given no information about what they may represent, and so could themselves be linear combinations of the "real" variables, or simply be highly dependent.
4. Random forest, SVM: Multiclass Support Vector Machines with Regression, and an ensemble method based on combination of these two models, enable us to predict the annual revenue of a restaurant.
5. We also experimented with a few algorithms ourself.

a. We tried ExtraTrees Classifier (A basic version of random forest) in place of k-means for Restaurant type substitution (explained later on)

b. We tried to use the Ridge model as an ensemble along with SVM and random forest, and as a standalone classifier.

ANALYSIS:

Pre-processing on City and Type greatly improves performance over baseline models.

Log transformation on revenue does not improve real test set performance although training set results are very promising.

Treating zero problem with PCA or KNN doesn't improve performance probably due to large misspecification errors of the treatment models that introduce more noise rather than clarity.

Apart from zero problem treatment, all solutions proposed in the paper together lead to least RMSE.

WHAT WE DID EXTRA:

We wanted to explore further than the methods proposed in the paper, and hence we looked up different models of classification. We also tried different kinds of preprocessing on the data types:

- For restaurant type such as T_MB, T_DT which were very rare we dropped it from the table and substituted those values with predicted type of **Extra Trees classifier** (a basic version of Random Forest)
- For the categorical and continuous problem we did a **one hot encoding** for "P" variables and took them as categorical. This greatly improved accuracy.
- We also **scaled all input features** between 0 and 1. This greatly improved accuracy.
- We tried different ways to taking the revenue apart from log(revenue), like **sqrt of revenue**, etc.

- A certain set of columns are either mostly all zero or all non-zero. We **added a feature** to mark this, storing the count of zero columns. This also greatly improved accuracy.
- We tried an alternative treatment for City problem, **replacing cities with their total counts**. This also greatly improved accuracy.

We also looked at the ridge model and tested it as an ensemble with SVM and random forest, and as a standalone classifier. Ridge as a standalone **gave us even better results than the methods mentioned in the paper and a Kaggle rank of 7. It consistently outperformed the other models with the corresponding RMSE score of 1,750,100.**

RESULTS:

P-Variable Treatment	Date Processing	City Problem Treatment	Type Treatment	Models Used	Ensemble Weighting	Revenue Treatment	Submission RMSE
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	Random Forest	None	None	2213632
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	Random Forest	None	Log Revenue	2007537
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	Random Forest	None	Sqrt Revenue	1947620
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	SVM	None	None	2148238
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	SVM	None	Log Revenue	1946873
Continuous	Number of days opened,	Replace with city	Extra Trees	SVM	None	Sqrt Revenue	1832463

	Year, Month	counts	Classifier				
Continuous	Number of days opened, Year, Month	K Means (K=20)	kNN Treatment (K=5)	Random Forest, SVM	Weighted	None	2054843
Continuous	Number of days opened, Year, Month	K Means (K=20)	kNN Treatment (K=5)	Random Forest, SVM	Weighted	Log Revenue	2023234
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	Ridge model	None	Log Treatment	1899332
Categorical	Log of Number of days opened	Replace with city counts	Extra Trees Classifier	Ridge model	None	Sqrt Treatment	1750100
Categorical	Log of Number of days opened	K Means (K=20)	Extra Trees Classifier	Ridge model	None	Sqrt Treatment	1793422
Categorical	Log of Number of days opened	Replace with city counts	Extra Trees Classifier	Ridge model, SVM, Random Forest	Weighted	Sqrt Treatment	1854353
Categorical	Log of Number of days opened	Replace with city counts	Extra Trees Classifier	Ridge model, SVM	Weighted	Sqrt Treatment	1784353
Categorical	Log of Number of days opened	Replace with city counts	Extra Trees Classifier	Ridge model, Random Forest	Weighted	Sqrt Treatment	1810323
Continuous	Number of days opened, Year, Month	Replace with city counts	Extra Trees Classifier	Ridge	None	None	1939160

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.

CONCLUSIONS:

When training data is small in size, **the simplest model often gives the best results.** In our case, compared to SVM and Random Forest the ridge model gave us the best results and heavily reduced our RMSE values.