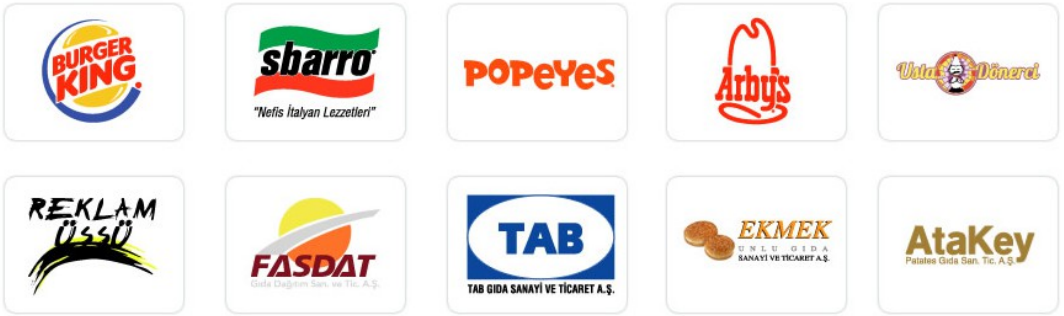
Restaurant Revenue Prediction



Objective

- **Predict annual restaurant sales based on objective measurements**

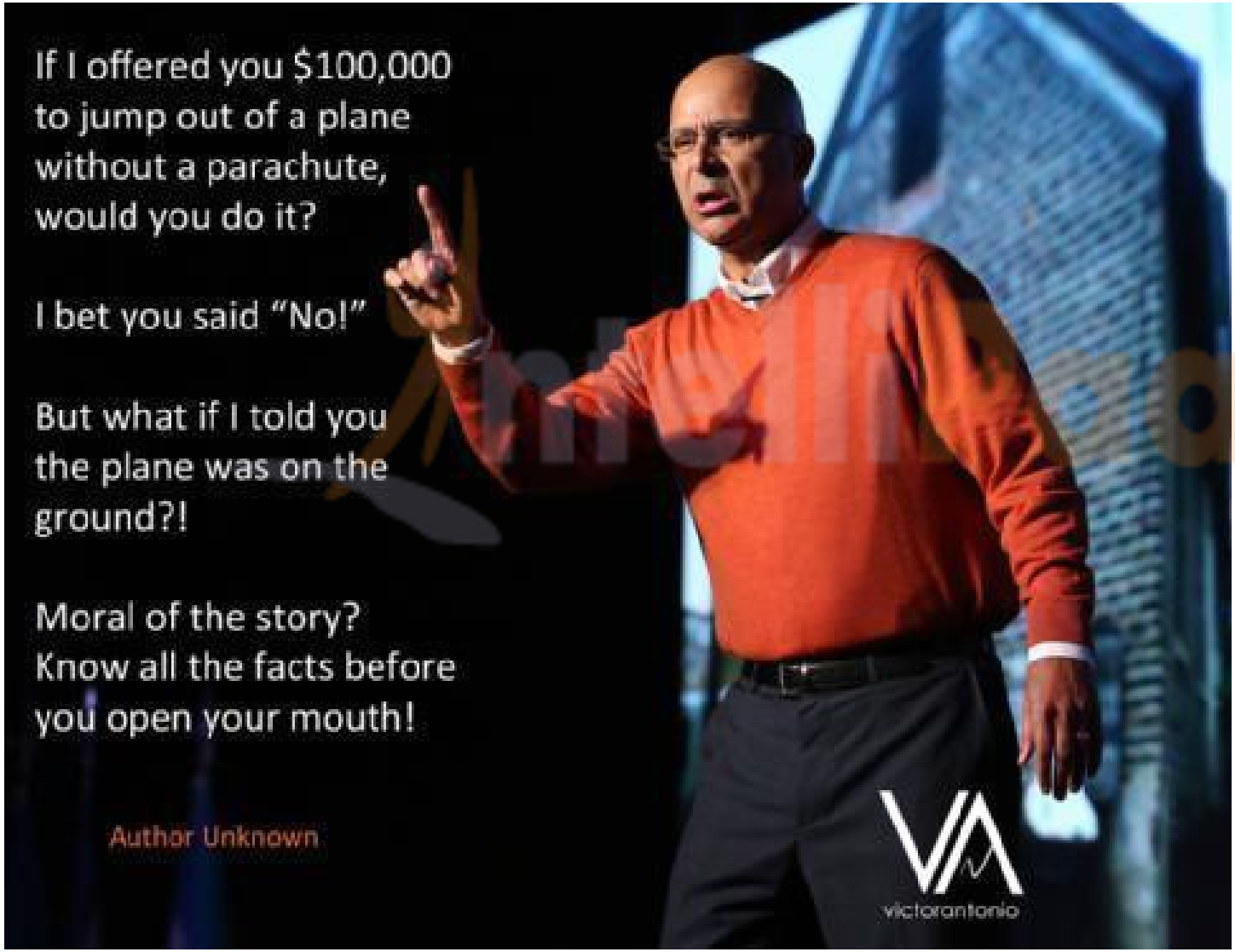
With over 1,200 quick service restaurants across the globe, TFI is the company behind some of the world's most well-known brands: Burger King, Sbarro, Popeyes, Usta Donerci, and Arby's. They employ over 20,000 people in Europe and Asia and make significant daily investments in developing new restaurant sites.

Right now, deciding when and where to open new restaurants is largely a subjective process based on the personal judgement and experience of development teams. This subjective data is difficult to accurately extrapolate across geographies and cultures.

New restaurant sites take large investments of time and capital to get up and running. When the wrong location for a restaurant brand is chosen, the site closes within 18 months and operating losses are incurred.

Finding a mathematical model to increase the effectiveness of investments in new restaurant sites would allow TFI to invest more in other important business areas, like sustainability, innovation, and training for new employees. Using demographic, real estate, and commercial data, this competition challenges you to predict the annual restaurant sales of 100,000 regional locations.

Know all the information before model it



Data Overview

- TFI has provided a dataset with 137 restaurants in the training set, and a test set of 100000 restaurants.
- The data columns include the open date, location, city type, and three categories of obfuscated data: Demographic data, Real estate data, and Commercial data.
- The revenue column indicates a (transformed) revenue of the restaurant in a given year and is the target of predictive analysis.

Data Fields

- **Id** : Restaurant id.
- **Open Date** : opening date for a restaurant
- **City :** City that the restaurant is in. Note that there are unicode in the names.
- **City Group**: Type of the city. Big cities, or Other.
- **Type**: Type of the restaurant. FC: Food Court, IL: Inline, DT: Drive Thru, MB: Mobile
- **P1, P2 - P37**: There are three categories of these obfuscated data. **Demographic data** are gathered from third party providers with GIS systems. These include population in any given area, age and gender distribution, development scales. **Real estate data** mainly relate to the m2 of the location, front facade of the location, car park availability. **Commercial data** mainly include the existence of points of interest including schools, banks, other QSR operators.
- **Revenue**: The revenue column indicates a (transformed) revenue of the restaurant in a given year and is the target of predictive analysis. Please note that the values are transformed so they don't mean real dollar values.

File Descriptions

- **train.csv(137 observations)** - the training set. Use this dataset for training your model.
- **test.csv(1 lakh observations)** - the test set. To deter manual "guess" predictions, Kaggle has supplemented the test set with additional "ignored" data. These are not counted in the scoring.
- **sampleSubmission.csv** - a sample submission file in the correct format

Links

http://
www.kaggle.com/c/restaurant-revenue-prediction

Data download link:

http://
www.kaggle.com/c/restaurant-revenue-prediction/data

EVALUATION USING RMSE(Root mean squared error)

- The use of RMSE is very common and it makes an excellent general purpose error metric for numerical predictions.
- Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors.

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

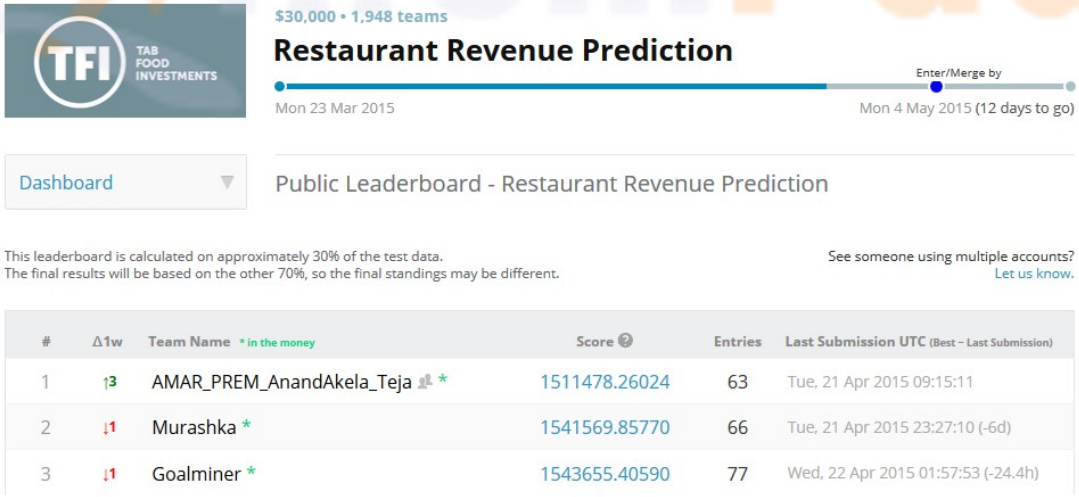
  - where y hat is the predicted value and y is the original value.
- Ex: RMSE= 100, Average revenue is 10000. which implies that the predict revenue is about 99% close to the actual revenue on an average sense.

Feature Engineering / Selection

- **Data Preprocessing and Transformation**
  - Perform data cleaning and preprocessing to remove outliers.
  - The data was partitioned into training(70%) and testing(30%).
  - Do log transformation on both dependent and independent variables.
- **Feature Engineering / Extraction**
  - Derived parameters using domain knowledge
  - Create row wise Mean, Stdev, min, Max, Count of non-zeros etc..
  - Using PCA(principal component analysis)
- **Feature Selection**
  - In this method the variables that have high correlation with dependent variable was chosen.
  - Wrapper algorithm
  - https://topepo.github.io/caret/index.html

Model vs Accuracy

- Different Models built and tested
  - Regression Model
  - Random Forest Model based on Feature Engineering
  - Random Forest Model based on Feature Selection
- Among these models "Random Forest based on Feature Selection" was chosen because of its good RMSE.



Random Forest Books

- "The Elements of Statistical Learning", Random forests are introduced in Chapter 15. If you're not familiar with regression trees you may also want to have a look at Chapter 9.2 as well where tree based methods are discussed.

http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

- **An Introduction to Statistical Learning with Applications in R**

http://www-bcf.usc.edu/~gareth/ISL/getbook.html

Alternatives

- glmnet: http://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html
- liblinear: http://www.csie.ntu.edu.tw/~cjlin/liblinear/
- caret is an interface package with several handy functions, more statistically oriented:

http://topepo.github.io/caret/index.html

- Model list supported:

http://topepo.github.io/caret/modelList.html

Advance Analytics

- Ensemble good models to get better accuracy.

  (RF+SVM+ Neural Nets etc..)

SVM(support vector machine) r code

- https://github.com/ujwlkarn/Restaurant-Revenue-Prediction/blob/master/Beat_the_Benchmark.R