Sales Prediction for Two Different Classes of Datasets

Final Project Report

by

Khushboo Agarwal (163050066)
Richa Verma (163050051)
Shifali Sonkar (163050083)
Kajal Gupta (163050031)

CS 725: Foundation Of Machine Learning

Computer Science & Engineering
Indian Institute of Technology, Bombay
Mumbai 400 076

Contents

1 Project Description

A crucial part of any business problem is the estimation of its future sales. There must be an idea of what money is it going to make in the coming time. Not only it helps with financing, but also in setting effective goals to stay longer in the business.

The problem of Sales Prediction can be framed as a learning problem where the objective is to develop a model and a set of preprocess procedures to accurately predict the revenues collected in a given time frame. The underlying datasets are diverse in that each has a different style of operation.

Our project purpose is to find a mathematical model to measure the effectiveness of investments in different business. Using demographic, real estate, and commercial data, we have to predict the future sales.

Since sales datasets can belong to different categories, under this project, we'll consider the the two different classes of datasets: Simple multivariate and Spatial datasets. We wish to deduce an appropriate generalized model that can be applied to any problem of sales prediction in future that belong to the above set of classes.

2 Project Approach

In simple terms, we need to find a Regression model to describe our problem, which will be used to predict the sales for any kind of business.

To support our different classes of datasets mentioned above, we are planning to use different tools like Random Forests, Support Vector Machines, K-Nearest Neighbour, Neural networks etc. We have evaluated our model by comparing the performance score on Kaggle.

For simple multivariate dataset, we can use different models like Random Forest, SVM, Naive Bayes etc. Whereas, K-Nearest Neighbour, Maximum Entropy, SVM etc can be used for spatial datasets. We plan to test different models for each class of datasets and finally come up with the one providing maximum accuracy for that particular class.

Programming Language: Python

3 SIMPLE MULTIVARIATE:

Restaurant Revenue Prediction

3.1 Problem Statement

The problem was to predict the annual restaurant sales of 100,000 regional locations using demographic, real estate, and commercial data.

3.2 Dataset Description

We have downloaded our datasets from Kaggle. Our training and tests data sets contains the following fields:

Restaurant Revenue Prediction

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 Through, 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.

3.3 Feature Engineering

Removed Restaurant 16 :Restaurant 16 has nearly identical values in all attributes with Id=38 and Id=85 .

City, City Group, Type Removed: These variables were loosely correlated and not very relevant for the prediction.

Taken selected values of Revenue: We have included only those values of Revenue attribute which are greater than 16000000.

Removed obfuscated attributes:We have removed attributes with high frequency zeros.

3.4 Regression Model

We have used 3 different models:

SVM: Support Vector Machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.We have used svm regression but this model is not very useful for our data-set.

Gradient boosting regression:Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Prediction of Revenue from this model is better than SVM.

Adaboosting:AdaBoost model is best used to boost the performance of decision trees on binary classification problems and for Regression.Prediction from this gave best result.

3.5 Observations

The below graph was observed before preprocessing, we can observe outliers in the graph (fig:1) clearly.

We trained our model using SVM, Gradient boosting regression and Adaboosting. We got best results and minimum error after using Adaboosing and with some parameter tuning.

Result after applying Adaboosting ,SVM, Gradient boosting

Model	Score(RMSE)
AdaBoosting	191361.5
SVM	2150340.810
Gradient Boosting	1998215.64

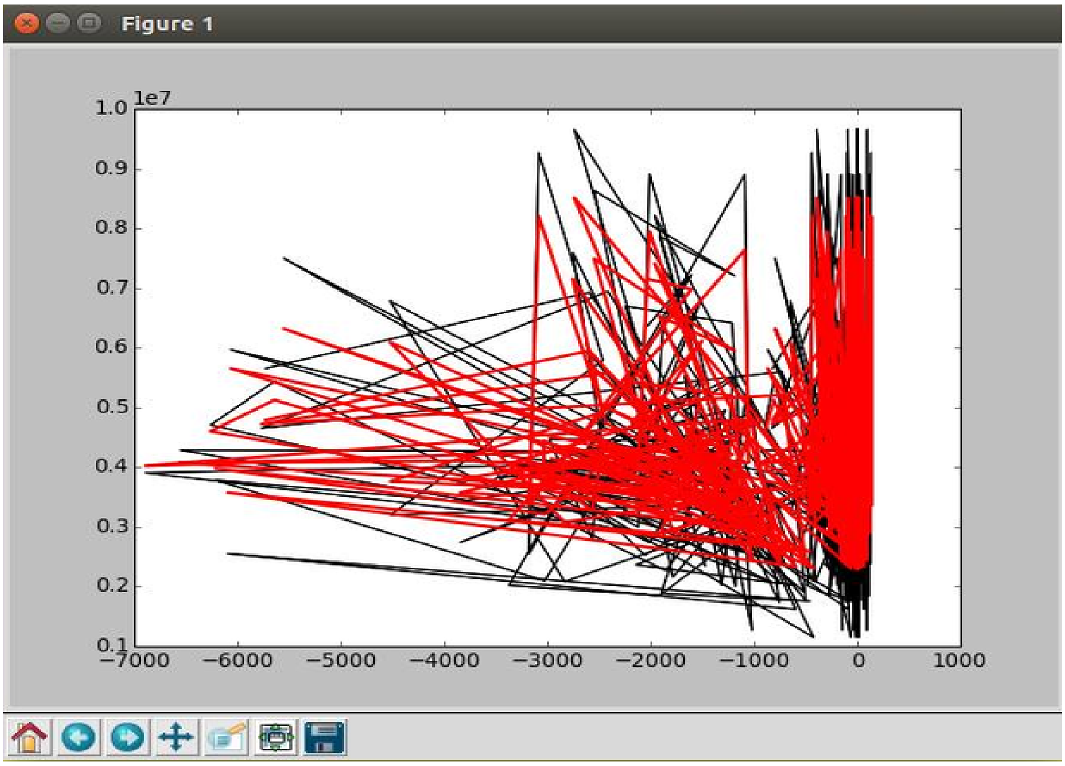

Figure 1: Unprocessed data without outlier removal

4 SPATIAL:

Sales in Stormy Weather

4.1 Problem Statement

The problem was to accurately predict the sales of 111 potentially weather-sensitive products (like umbrellas, bread, and milk) around the time of major weather events at 45 of their retail locations.

4.2 Dataset Description

We have downloaded our datasets from Kaggle. The dataset was spread over 3 different files as follows:
weather.csv : It contained the weather conditions for different weather stations.
key.csv : It contained the station number to store number mapping.
train.csv : It contained the sales for different stores on some particular dates.

test.csv : Here we need to preduct the sales for different stores on some particular dates.

Our training and tests data sets contains the following fields:
Sales in Stormy Weather

date: the day of sales or weather

store_nbr: an id representing one of the 45 stores

station_nbr: an id representing one of 20 weather stations

item_nbr: an id representing one of the 111 products

units: the quantity sold of an item on a given day

id: a triplet representing a store_nbr, item_nbr, and date. Form the id by concatenating these (in that order) with an underscore. E.g. "2_1_2013-04-01" represents store 2, item 1, sold on 2013-04-01.

4.3 Feature Engineering:

Removal of Store-Item combination with zero sales: We actually noticed that sales value for most of the stores item number combination were zero. So, we checked mean unit for every store and item numbers. If mean was zero, we just removed the store and item number combination. It was assumed that the store and item number which never showed up any sales in past, will not show in future also.

Weather flag: We added two flags depart and precipitation flag for which average departure should be greater than 8 for being 1, less than -8 for being -1 and 0 otherwise.For average precipitation > 0.2, precipitation flag was made 1, 0 otherwise. This is done as the weather conditions largely influenced the sales for an item at a specific location.

Holidays: Using a holiday calendar downloaded from Internet, we checked whether each date was a holiday or not.

Scaling the number of units: Units were logarithmically transformed in order to scale them down.

Date to days conversion: All dates were converted into number of days by taking its difference from a reference date. Like here, it was 01-01-2012.

4.4 Regression Model

We have tried 3 different models for our use:
K-Nearest-Neighbour The k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for regression. For the regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

We have used KNN on our preprocessed data. The maximum neighbour size was varied and tested for different values. We got the optimum result on neighbourhood size of 10.
Gradient Boosting Gradient boosting is a machine learning technique for regression, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

We varied the learning rate and maximum depth. The optimal value was obtained at .

Support Vector Regression Support Vector Regression (SVR) are supervised learning models with associated learning algorithms. It depends only on a subset of the training data, because the cost function for building the model does not care about training points that lie beyond the margin.

We trained and tested our model for SVR rbf kernel, C = , gamma = .

4.5 Observation

We trained and tested our dataset using Gradient Boosting with learning rate 0.08(obtained through hit-and-trial). We varied the max-depth of tree and the error obtained was as follows:

Maximum Depth	Error(RMSLE)
12	0.15793
9	0.17022
7	0.21039
5	0.3028
1	0.3892

Similarly for KNN, by varying the max-neighbour size, the output obtained was as follows:

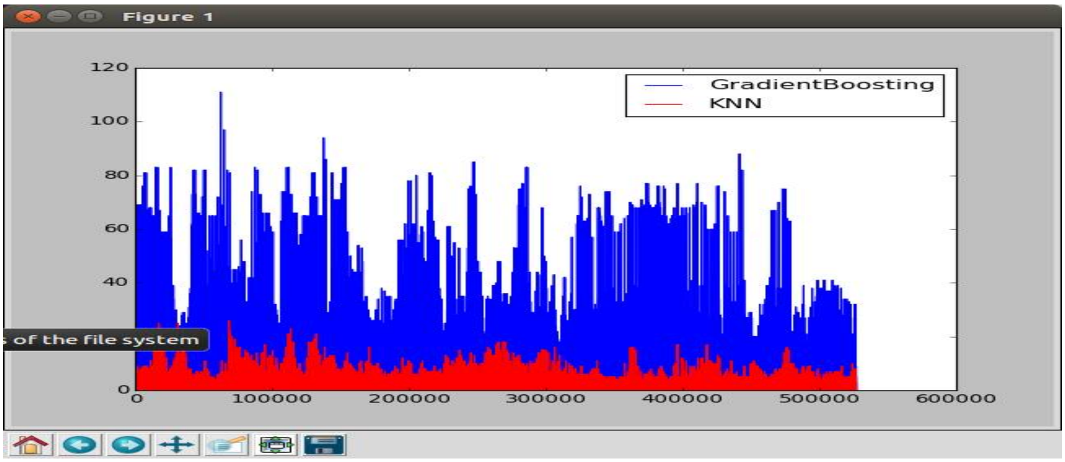

Figure 2: Gradient Boosting vs KNN

Maximum Neighbour	Error(RMSLE)
8	0.4031
10	0.3901
15	0.4101

We can simply see from the observations that the performance obtained through Gradient Boosting was comparatively better than KNN or SVR for our dataset. We can see the difference in output predictions of the two models(Gradient Boosting and KNN) using the graph shown in figure 2.

5 Conclusion

Our project objective was to find mathematical models to measure the effectiveness of investments in different business setups. We used 2 different datasets, applied feature engineering on them. And, then finally using diffemt machine learning tools, we trained a model to predict sales unit for unseen data.

We observed that Boosted Tree Regression performed comparatively better on our both set of datasets.

6 References

1. Prof. Nataasha Raul, Yash Shah, Mehul Devganiya. **"Restaurant Revenue Prediction using Machine Learning"** International Journal of Engineering And Science Vol.6, Issue 4 (April 2016)

2. Thiesing, Frank M., and Oliver Vornberger. **"Sales forecasting using neural networks."** Neural Networks, 1997., International Conference on. Vol. 4. IEEE, 1997.

3. Sharma, Rashmi, and Ashok K. Sinha. **"Sales Forecast of an Automobile Industry."** International Journal of Computer Applications 53.12 (2012).

4. Giering, Michael. **"Retail sales prediction and item recommendations using customer demographics at store level."** ACM SIGKDD Explorations Newsletter 10.2 (2008): 84-89.