Restaurant Revenue Prediction using Machine Learning

Srikanth Siddamsetty[1], Rohith Reddy Vangala[2], Lokesh Reddy[3], Praneeth Reddy Vattipally[4]

Abstract - *Personal judgment and experience will influence the location and release date of a restaurant. It can be challenging to extrapolate subjective data across cultures and geographical areas. A supervised learning algorithm can construct complicated representations of complex variables using simple inputs such as the opening date, city and type of the restaurant (Food Court, Inline, Drive-Thru, Mobile), Demographic data (population in any given area, age, and gender distribution, development scales), Real estate data (front facade of the location, car park availability), and points of interest including schools, banks. Machine learning concepts such as catboost and random forests can enable food chains to predict the annual revenue of a new restaurant, which can help them determine the viability of a news outlet.*

Key Words: Machine Learning, Random Forest, Catboost, Restaurant, Revenue, Prediction.

1.INTRODUCTION

The establishment of new restaurant outlets incurs substantial time and capital expenditures. When the news outlet fails to break even, the site closes within a short time after operating losses are incurred [2]. Finding an algorithmic model to increase the return on investments in new restaurant sites would facilitate businesses to direct their investments in other critical business areas, like innovation and training for new employees. We propose an automated method for determining the task environment for a new restaurant by using concepts of Support Vector Machines, Gaussian Naive Bayes, and Random Forest. Predicting the annual revenue of a restaurant will help food chains determine the feasibility of opening a new outlet. By utilizing Machine Learning to predict the annual revenue of restaurants, restaurants can make better decisions about opening new outlets [4]. The proposal aims to find an algorithmic model that will help increase the efficiency of investments in new restaurant sites. The proposal seeks to predict the revenue of new outlets of existing restaurant chains as one of its main features. Analytical prediction of data has proven more effective than human judgment. More importantly, it is capable of analysing and comparing multiple new sites. The procedure eliminates human errors and allows operations more rapidly than before. A dataset with 37 obfuscated parameters will be employed to train the algorithm, no more [5].

Establishing new restaurant outlets involves major time and capital expenditures. Whenever a new outlet cannot handle the influx of customers, the site closes in a short time, causing extensive operating losses. It would be advantageous if algorithmic models existed to increase return on investment in new restaurant sites. This would allow businesses to send theirinvestment dollars to other important areas, such as innovation and training new employees.

Machine learning helps predict restaurant revenue to help restaurants make more informed decisions about launching new outlets. Efficiencies of new restaurant investments would improve by creating a model using algorithms. Predicting the profits of new restaurant outlets for existing restaurant chains is an essential function of the proposed application. Predictions based on data analysis have proven more reliable than human judgment. Moreover, it allows for a comparison of multiple new sites.

2. METHODOLOGY

2.1 Catboost Algorithm

It provides a gradient boosting framework that attempts to solve for categorical features using a permutation-driven alternative compared to the classical algorithm. Yandex has developed an open-source software library known as CatBoost. It works on Linux, Windows, macOS, and is available in Python, R, and models built using catboost can be used for predictions in C++, Java, C#, Rust, Core ML, ONNX, and PMML. The source code is available on GitHub under the Apache License.

There are several reasons why CatBoost has gained popularity over other gradient boosting algorithms:

- Ordered Boosting to overcome overfitting
- Native handling of categorical features
- Using Oblivious Trees or Symmetric Trees for faster execution

CatBoost can offer a wide range of solutions to a wide range of data problems thanks to its integration of diverse data types by numerous businesses.

There is a growing community at CatBoost that seeks feedback and contributions from the users. The community is available via Slack, Telegram (with English and Russian versions), and Stack Overflow. You can report bugs via the GitHub page if you discover one

2.2 Random Forest Algorithm

In ensemble learning, random forests or random decision forests are used for classification, regression, and other tasks based on the construction of many decision trees at training time. Random forests are used for classification tasks, where the output is the class selected by the majority of trees. Regression tasks return the mean or average prediction of

individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.:587–588 Random forests generally outperform decision trees, but their accuracy is lower than gradient-boosted trees. Their performance can, however, be affected by data characteristics. Tin Kam Ho developed the first random decision-tree algorithm in 1995 using the random subspace method, which in his formulation implements the "stochastic discrimination" approach to the classification proposed by Eugene Kleinberg. This algorithm was improved by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark in 2006 (as of 2019, owned by Minitab, Inc.). The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance. Random forests are frequently used as "blackbox" models in businesses, as they generate reasonable predictions across a wide range of data while requiring little configuration.

3. IMPLEMENTATION

Here we use Machine Learning algorithms to predict the revenue of the restaurants. The algorithms used are catboost and random forest. The process flow of this project is represented in Fig 3.3. Initially the data is collected and studied. Secondly the raw data is structured and divided into train data and test data. Now the data is provided to the algorithms and allowed to train the data. First it is checked that the data is valid or not. If data is valid then the algorithms get trained separately. Here we are implementing this project using two algorithms because to know which algorithm is predicting with least mean absolute error and least root mean square error. So, the two algorithms are trained accordingly for different parameters of input. Then next it comes to the test data where the data is provided to predict the revenue of the restaurants based on the parameters provided. This data is compared with data used for training algorithm and then the errors are calculated. The less the error is the good the efficiency and accurate the prediction. So, the restaurants revenue is predicted based on the trained data or the historic data.

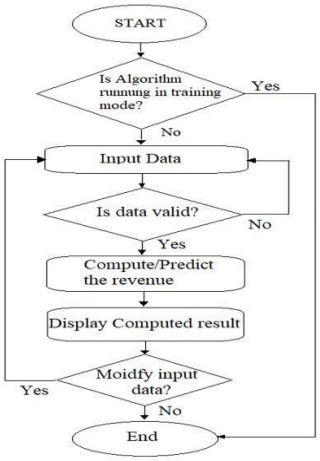

Fig- 1: Data Flow Model of the whole process

4. DATASET

The dataset contains parameters (as in Fig 2): 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

	Id	Open Date	City	City Group	Type	P1	P2	P3	P4	P5	...	P29	P30	P31	P32	P33	P34	P35	P36	P37	revenue
0	0	07/17/1999	İstanbul	Big Cities	IL	4	5.0	4.0	4.0	2	...	3.0	5	3	4	5	5	4	3	4	5653753.0
1	1	02/14/2008	Ankara	Big Cities	FC	4	5.0	4.0	4.0	1	...	3.0	0	0	0	0	0	0	0	0	6923131.0
2	2	03/09/2013	Diyarbakır	Other	IL	2	4.0	2.0	5.0	2	...	3.0	0	0	0	0	0	0	0	0	2055379.0
3	3	02/02/2012	Tokat	Other	IL	6	4.5	6.0	6.0	4	...	7.5	25	12	10	6	18	12	12	6	2675511.0
4	4	05/09/2009	Gaziantep	Other	IL	3	4.0	3.0	4.0	2	...	3.0	5	1	3	2	3	4	3	3	4316715.0

Fig. 2: Snapshot of the dataset

5. RESULT

After the data is collected from different sources then the collected is to be studies and statistical analysis is to be done. The different analysis is represented in Fig. 3 and Fig.4. The predictions for different restaurants are as follows

From Fig. 3 represents that the restaurants from big cities are more in number than the other cities category. So finding out the number of restaurants in different categories helps to train the algorithm effectively and to make it understand better.

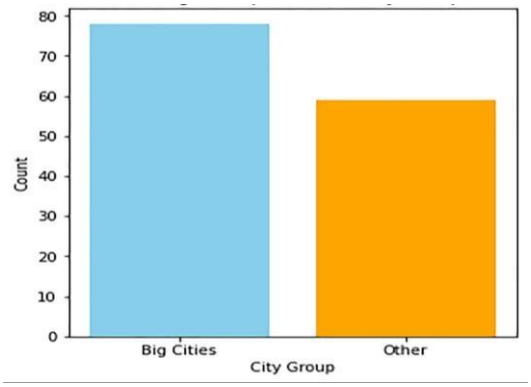

Fig-3: Statistical representation of Restaurants in different cities+

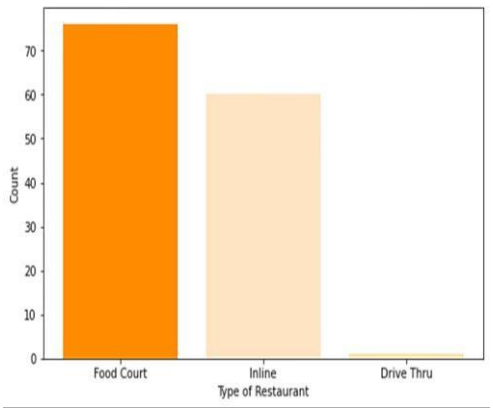

Fig-4: Statistical Representation of type of Restaurants

From Fig. 4 is the representation of the number of restaurants in different category like Food court Inline Drive thru. This analysis makes us and the algorithm understand easy about the data collected and the predictions will be accurate.

Two machine learning algorithms are used named as Catboost and Random forest for predicting the revenue from the input parameters. And the result of the Catboost algorithm that is the revenue prediction for different restaurants are predicted and the below Fig. 5 represents the predictions of the revenue.

Fig-5: Revenue Predicted using Catboost algorithm

The predictions of the catboost algorithm are more accurate than the other algorithms which in turn tells us that the revenue is predicted correctly. The predictions in above result are for many restaurants which is predicted from the previous history data.

The next model or algorithm is the random forest algorithm using which the revenue of different restaurants are predicted. Initially this model is trained by providing with the history data of different restaurants and making it understand and then providing the test data to predict the revenue and as a result the below Fig. 6 represents the predicted values of the different restaurants having unique Id's.

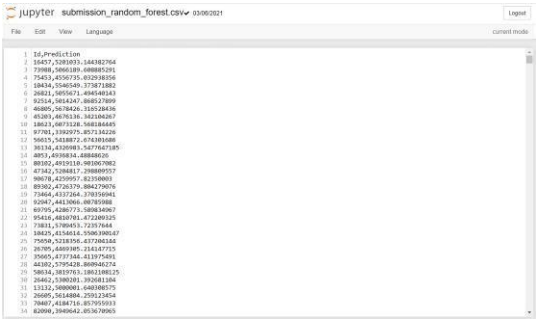

Fig-6: Revenue Predicted using Random Forest algorithm

Table -1: Comparison between proposed solution and base paper

Parameter	Proposed Solution	Base Paper
Root Mean Square Error	0.914	0.6194
Algorithms	Catboost and Random Forest	Bayesian Linear Regression and Boosted Decision Tree Regression

6. CONCLUSION

This can be emulated as a way of developing a forecasting system for restaurant revenues. We predicted annual restaurant sales by using Random Forests and CatBoost. It is

through this approach that a human judgment aid can be provided and losses in food chains can be minimized, providing a reference.

REFERENCES

[1] SauptikDhar, Vladimir Cherkassky, "Vizualization and Interpretation of SVM Classifiers", Wiley Interdisciplinary Reviews

[2] Geoffrey Hinton, Sam Roweis, "Stochastic Neighbor Embedding",

University of Toronto

[3] "Restaurant Opportunities in India: Trends and Opportunities", http://www.hvs.com/Content/1336.pdf , 2004

[4] Wen-Chyuan Chiang, Jason C.H. Chen, XiaojingXu, "An overview of research on revenue management: current issues and future research, International Journal of Revenue Management", Vol. 1, 2007

[5] "Dataset: Restaurant Revenue Prediction", https://www.kaggle.com/c/restaurant-revenue-prediction