IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 9, Issue 12, December 2020

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

Analytical Tool and Sales Forecasting Model

Ramya D[1], Vasantha S[2], Vishnuvardhan Y[3]

PG Student, Department of Computer Applications, Sir MVIT, Bangaluru, Karnataka, India[1]

Assistant Professor, Department of Computer Applications, Sir MVIT, Bangaluru, Karnataka, India[2]

Project Manager, Exposys Data Labs, Yelhanka, Bangaluru, Karnataka, India[3]

ABSTRACT: Supervised learning algorithm will construct complex features using simple features such as opening date for a restaurant, city that the restaurant is in, type of the restaurant, demographic data (population in any given area, age and gender distribution, development scales). Applying concepts of machine learning such as support vector machines, random forest, linear regression, Lasso regression and Logistic regression, on these parameters, it will predict the annual revenue of a new restaurant which would help food chains to determine the feasibility of a newoutlet.

KEYWORDS:Random forest; Lasso regression; linear regression; Support vector.

I. INTRODUCTION

New restaurant outlets cause tremendous time and capital ventures to set up. At the point when the new outlet neglects to make back the investment, the site closes inside a brief time frame and working misfortunes are caused. Finding an algorithmic model to expand the return on interests in new café locales would encourage organizations to coordinate their interests in other significant business zones, similar to advancement, and preparing for new workers. The issue be characterized as plan an computerized way to deal choose the assignment condition for new restaurant by applying ideas of Support Vector Machines, Gaussian Naïve Bayes and Random Forest on specific parameters, it will foresee the yearly income of a new restaurant outlet which would help evolved ways of life to decide its practically. The essential goal of Restaurant income prediction utilizing Machine Learning is assist Restaurants with settling on increasingly educated and ideal choice about opening new outlets. It accepts to locate an algorithmic model to build the viability of interests in new restaurant destinations. Probably the greatest component of the proposed Applications is that it plans to see the income of new outlets of existing cafenetworks.

II. LITERATURE REVIEW

Machine learning algorithms were studied for predicting the annual revenue of new restaurants. Research papers on Support Vector Machines and Random Forest were referred. A research study on restaurant opportunities in India was read for better understanding.

- Support VectorMachine

At first approximation what SVMs do is to find a separating line (or hyper plane) between data of two classes. SVM is an algorithm that takes the data as an input and outputs a line that separates those classes if possible. According to the SVM algorithm that can be find the points closest to the line from both the classes. These points are called support vectors. Now, we compute the distance between the line and the support vectors. This distance is called the margin. The goal is to maximize the margin.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

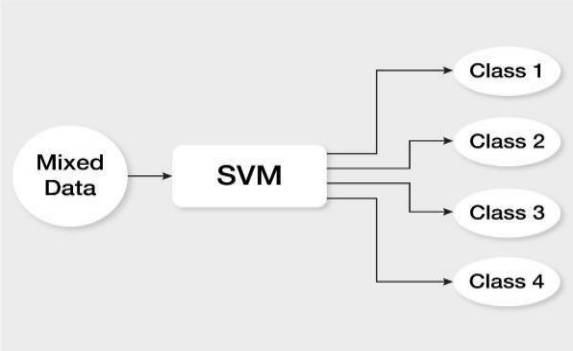

Fig 1: classification of SVM

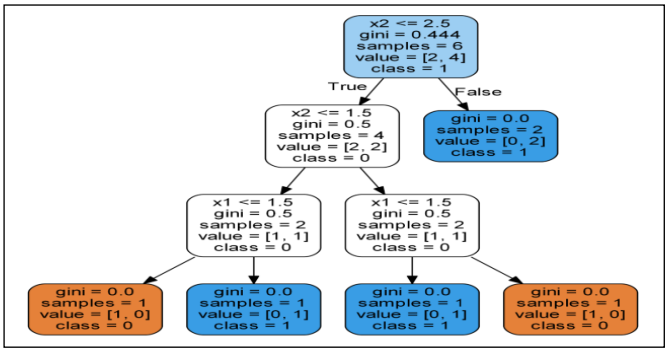

Fig 2: Decision tree classification

As the support vector classifier works by putting data points, above and below the classifying hyper plane there is no probabilistic explanation for the classification.

➢RandomForest

The technical details of a decision tree are in how the questions about the data are formed. In the CART algorithm, a decision tree is built by determining the questions (called splits of nodes) that, when answered, lead to the greatest reduction in Gini Impurity. What this means is the decision tree tries to form nodes containing a high proportion of samples (data points) from a single class by finding values in the features that cleanly divide the data into classes.

Our data only has two features (predictor variables), x1 and x2 with 6 data points — samples — divided into 2 different labels. Although this problem is simple, it's not linearly separable, which means that we can't draw a single straight line through the data to classify the points. however draw a series of straight lines that divide the data points into boxes, which we'll call nodes. In fact, this is what a decision tree does during training. Effectively, a decision tree is a non-linear model built by constructing many linearboundaries.

➢ Random Forest Classifier

A random forest is a meta estimator that fits a number ofdecision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

In the above figure shown the decision tree classification that can explained as below:

- Question asked about the data based on a value of a feature. Each question has either a True or False answer that splits the node. Based on the answer to the question, a data point moves down thetree.
- gini: The Gini Impurity of the node. The average weighted Gini Impurity decreases as we move down thetree.
- samples: The number of observations in thenode.
- value: The number of samples in each class. For example, the top node has 2 samples in class 0 and 4samples in class1.
- class: The majority classification for points in the node. In the case of leaf nodes, this is the prediction for all samples in the node.

➢ Random Forest Regressor:

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning. Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees.

In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs. A random forest is a meta estimator that fits a number of classifying of decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over- fitting.The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build eachtree.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

III. PROPOSED SYSTEM

This project considers and examines various parameters from the given data sets. The outcome of the analysis will display improved investments on the restaurant business. The objective is to assist people wanting to make investments in the restaurant business Prediction of Revenue for all the restaurants in the training data set using different predictive models.

- GBM(Gradient boosting machines)produces a prediction model in the form of an ensemble of weak prediction models which typically are decision trees. This model is used to handle variety of lossfunction.
- LR(Lasso Regression) is an absolute shrinkage and selection method used for analyzing the performance for both variable selection and regularization to enhance the prediction accuracy and interpretability of the statistical model itgenerates.
- Visualizing all the obtained algorithms and finding the differences and spotting out the best for userexperience.

➢ *DATASET*

Dataset consist of the following fields:

- Id : Restaurant id.
- Open Date : opening date for arestaurant
- City : City that the restaurant isin.
- City Group: Type of the city. Big cities, orOther.
- Type: Type of the restaurant.
- FC: FoodCourt.
- IL:In-Line.
- DT: Drivethru.
- MB:Mobile.

P1 - P37: These are the p-variables and are a measure of the following:-

- Demographic data:Population, age, gender distribution, developmentscale
- Real Estate Data: M2 of the location, front façade, car parking etc.
- Commercial data: schools, banksetc.
- REVENUE: The revenue column indicates a (transformed) revenue of the restaurant in a given year and is the target of predictiveanalysis.

➢ *Revenue Distribution*

Plotting a simple histogram of the revenue field show that it is not normally distributed but rather has a long tail toward the right. One simple solution to this problem is to take the ln of the revenue field and obtain something that is more similar to a normal distribution.

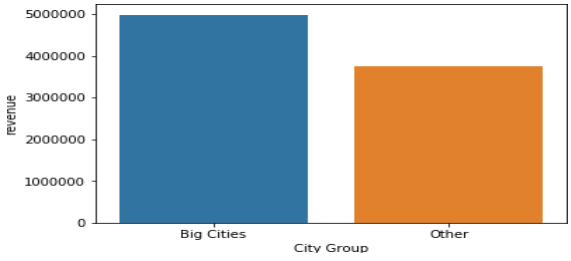

Fig 3: Plotting of revenue

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

- *Advantages of proposedsystem*

  - Random Forest Works the best on our pre-processeddata.
  - Gradient Boosting Machine is working much better thanSVM.
  - Getting the more accuracy then existingmethod.
  - Visualization is very helpful forusers.

.

IV. IMPLEMENTATION

- **Pre-Processing:**Before start to build a spatial model, it's essential toconduct a data preprocessing phase in order to: fill-in missing data, convert different data types, and standardize and normalize the features ifnecessary.

- **Analysis:**LASSO is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model.

- **Featureselection:**Classes in the sklearn. feature_selection module can be used for feature selection/dimensionality reduction on sample sets, either to improve estimators' accuracy scores or to boost their performance on very high-dimensional datasets.

- **Train and Test Split for Random ForestClassifier:**Thetrainingsetcontainsaknownoutputandthemodellearnson this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on thissubset.

- **PCA:**The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset (or subset) in order to test our model's prediction on this subset.

  - **Predicting revenue:**

```
Score is calculated as:  0.8348005123469309
```

Random forests are promising, because they have desirable characteristics, such as running efficiently on large datasets and flexibility, having been used effectively for a variety of applications. A random forest is an ensemble of decision trees created using randomvariableselection and bootstrap aggregating (bagging). What this means is that first a group of decision trees are created. For each individual tree, a random sample with replacement of the training data is used for training. Also, at each node of the tree, the split is created by only looking at a random subset of the variables. The prediction is made by averaging the predictions of all the individual trees. Random forests can also provide an error estimate, called the out-of- bag error. This is computed by feeding the individual trees cases that they have not seen. The training data that was not included in the bootstrap sample for a given tree is fed through that tree, and a prediction is made.

The results of doing this for all trees are computed, and for each sample (row) an out of bag estimate is created by taking the mean of the results of all thetrees.

V. RESULT

Types of documents that can actually represent results include policies, plans, procedures, reports, specification etc. The term results can and should be contrasted with products and or services. Also see deliverable for moreinformation.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

||Volume 9, Issue 12, December 2020||

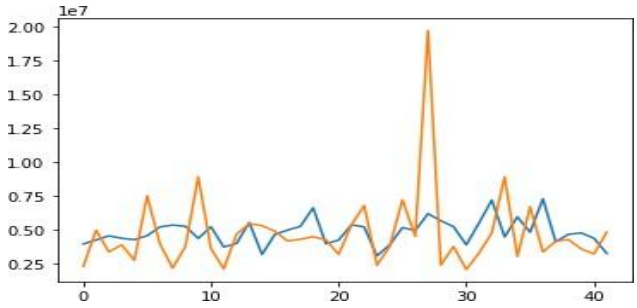

Fig 4: Predicting revenue using graph

The above Figure indicates Actual revenue (in orange, uppermost line) to Predicts the annual revenue as it produced a wider range of values whereas the predictions by Random forest (in blue, middle line).

VI. CONCLUSION

During the data analysis we observed the effects of PCA and RBF on our model. Also we had a chance to decide which technique to use for prediction. We choose random forest regression because it is more useful compared with other algorithms. The future revenues ofnew restaurant outlets can be analyzedusing Random forest and SVM were used to predict the annual revenue. In data preprocessing, we examined certain characteristics of the data set and provided potential fixes with PCA. Within Model selection, we focused primarily upon SVM and Random Forest implementations as they performed the best. Using this, a reference can be provided to aid human judgment and operational losses can be minimized for food chains.

REFERENCES

1. S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed.,R. M. Osgood, Jr., Ed. Berlin, Germany: Springer- Verlag,1998.
2. J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol.61.
3. S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, ―A novel ultrathin elevated channel low-temperature poly-Si TFT,‖ IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov.1999.
4. M.Wegmuller,J.P.vonderWeid,P.Oberson,andN.Gisin,―High resolution fiber distributedmeasurements with coherent OFDR,‖ in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.
5. R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, ―High-speed digital- to-RF converter,‖ U.S. Patent 5 668 842, Sept. 16,1997.
6. M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
7. A. Karnik, ―Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,‖ M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan.1999.
8. J. Padhye, V. Firoiu, and D. Towsley, ―A stochastic model of TCP Reno congestion avoidance and control,‖ Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02,1999.
9. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11,1997.
10. Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58.1 (1996):267-288.
11. Meinshausen, Nicolai, and Peter Bühlmann. "Stability selection." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72.4 (2010):417-473.
12. Kursa, Miron B., and Witold R. Rudnicki. "Feature selection with the Boruta package." J Stat Softw 36.11 (2010):1-13.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details