RECOMMENDER SYSTEM FOR FOOD STARTUP

Naga Raju M

1Assistant Professor, Department of Computer Science, GSS, GITAM (Deemed to be) University, Bengaluru-562163, Karnataka, India

Abstract: Automatic suggesting models have been available already for item buying, films, and amusements. Demand for one such system for food delivery startup is picking up now. Main aim of this research work is to come up with design of one such recommender system for food delivery startup makes it to sustain in the business on long run through better understanding as well as retention of potential customers automatically. Zomato reviews data are taken for analyzing using some selected algorithms of machine learning. From them the best predictor model is picked up for building the final proposed engine. Underlying methodology goes like this, it first knows and learns about customers using past and near past data (training data) and using that knowledge it tries to draw results of current data. Accuracy of this result is then observed and pruned, if not up to considered bench mark, by employing bias-variance tradeoff as well as considering more reliable characteristics. Once classifier design with required level of accuracy is obtained and it can then be developed. The focus of the research is to design a system that facilitates automatic recommendation for food startup to know and retain its good restaurants aggregation as well as potential users.

Keywords: intelligent systems, learning and prediction parts of data collection, basic and inferred characteristics, predictor.

I. INTRODUCTION

Digital feedback of users of food delivery startup, Zomoto, that are obtained from its site actually constitutes the dataset. These feedbacks are insisted to be in two valued form like or not like for the sake of uniformity in reviews from the users. They are collected as part of user’s views after dining the food.

Article is organized as 1) first part doing class label prediction to the best possible extent 2) second part focusing on increase in accuracy by taking extra and related parameters.

Usually in many food startups automatic proposing of food choices to the users based on their past preferences is missing. Aim of this research work is to produce such system for food startup. Dual valued user feedback, like yes or no, is preferred to have uniformity of the information. Descriptive nature of user feedback misses this uniformity factor and thus leads to clutter that raise complexity in the development of the system.

II. DATA PROCURING

Zomato startup’s reviews information is drawn from Kaggle website. Data collection is done till 14th October 2019. This collection contains 52000 records and 17 attributes for each record. Rows represent users and columns denote characteristics of food delivery outlet. Some of the attributes are homepage link, area, outlet title, order_by_web, reserve_buffet,
The following diagram depicts three parts of data set

![Diagram showing data set partition]

The model is designed in a manner that it first gets trained on past and near past data. Later tries to do prediction on the available test data. Precision of this guess is looked up. It is pruned to have high level of accuracy. Rightly functioning model is then tried to do good forecast for selected user. 5-point measurement system is then mapped two valued system in manner as favored-(4,5) and not favored-(1,2,3). Threshold of accuracy is taken as 72% [1] which means that the model performing guess up to 72% will be regarded as doing fine.

### III. TRAITS DESCRIPTION

Traits of data records [5] taken are described as below.

- **homepage**: Website of food startup.
- **area**: Place of food startup.
- **Outlet title**: Label of food startup. Few examples are Café, Coffee Day, Onesta, Just Bake, Kranti sweets and so on.
- **Order_by_web**: Provision to make booking through online. More than 30000 number of food points of Zomato facilitating this option. Around 19,000 are not doing so.
- **reserve_buffet**: Buffet booking in advance.
- **outlet_rating**: Ranking of food outlet.
- **number_polled**: Number of users choosing the outlet.
- **communication_info**: Outlet contact information.
- **eatery_class**: Different categories of eateries of the outlet.
- **item_cost and average_item_cost**: Price and mean price of eateries of the outlet.
fond_of_food: Love food at the outlet.

items_consumed: Different items available to users to have it.

approx_cost_of_dining: Mean cost of dining for group of people, which is calculated as total dining cost / number of people dining.

blend_cost and mean_value: Value of each blend of food and its mean value.

dish_card: Menu card of dishes served at the outlet.

data_list: Users feedback on food outlet after dining.

IV. PREDICTORS MODELLING

The columns multiple values are mapped into binary labels through the concept of binary encoding given in [8]. This encoding is done in a way as (4,5) / (1,2,3) to prefer / not_prefer values. Less influential columns are ignored due to reason that they will have less impact on the final model going to be obtained.

The dataset considered is basically tuned into the following three types of sets. Type 1 set considers direct characteristics. Type 2 set includes indirect characteristics. Type 3 set consists of combination of both of them.

Predictors of different algorithms under consideration are designed. It is first made to learn and used to do the prediction on training and testing data respectively. Outcome of them are observed. Algorithm producing better outcome is taken for final design of the predictor.

Pick RBFSVM, LSVM and LR algorithms of machine learning and construct heir predictors. The characteristics like average costs, rating, different combo offers, and food varieties having significant impact on the final result are taken. Along with them some of the indirect features having equal significance on final result are also considered. Few of them are total number of orders per day and popular outlet. The values of them are actually obtained from average costs and rating attributes. In each of these models hyper plane separating these review data points clearly into groups is fitted. The fitting is then transformed into the best fit by calculating error and minimizing it to the substantial level. Predictor models are designed based on the best fit. Comparison of final results of three predictor models is done. The predictor model producing better result among them is then finally chosen. Recommender system based on that predictor model is going to be developed and deployed to the startup. In this whole process one observation is apparent. It is the consideration of type 3 dataset where combination of both features exists and that is actually yielding better results than that of individual feature existing in other two types of sets.

V. BRIEF DESCRIPTION OF THREE MACHINE LEARNING ALGORITHMS

Linear Regression: It uses least squares to fit a line to the data taken. Next determine R². Finally find out P-value for R². Objective of Linear Regression is to get best fit for the data under consideration. The procedure to arrive at best fit goes as follows. Assume to move all data on to Y-axis and get their mean. Draw a line moving through that mean value of Y-axis. Measure
distance of each data point from that line, called residual. Take the square of them to get positive values. Find sum of squares of residuals. Rotate the direction of line and find sum of squares of residuals. Repeat this until the rotation of line produces least sum of squares of residuals. That line serves as the best fit for the data under consideration. The non-zero slope of this fit specifies that independent variable in fact helps in making a guess of dependent variable. Next to see that how well this guess will be. It is done using R². As part of it first find out SS(mean), sum of squares around the mean=(data-mean)². Var(mean), variation around the mean=(data-mean)²/n, where n is the sample size. In other words, Var(mean)=SS(mean)/n, average sum of square around the mean. Find out SS and Var for the best fit in a similar manner. Observe that Var(fit) will be far less than Var(mean). This way R² confirms that residuals in the fit are significantly smaller which in turn says that the guess is far good in nature. The basic formula for R²=(Var(mean)-Var(fit))/Var(mean). Next consider to find out P value for R². P value is calculated with F distribution. The degree of freedom turns sum of squares into variances. In general P denotes number of parameters in each of mean and good fit of data under consideration. P also allows considering extra parameters and finding out their impact on the final result. This can be measured with the formula of F. The values of numerator and denominator of F for a good fit can be interpreted as, variation explained by extra parameters will be larger number whereas variation not explained by extra parameters will be smaller number. Thus R² quantifies the relationships between data which should as large as possible. Whereas P value measures the reliability of that relationship between data and it should be as small as possible.

Linear and Radial Basis Function (RBF) SVM: First job is to find threshold of the data points under consideration that classifies them into two groups clearly. Randomly choosing a threshold most of the times lead to misclassifications. Instead take edge data points of each group and take midpoint of them as threshold. The distance between these edge data points of each group and threshold is called maximal margin. This is venerable to outliers and leads to misclassifications. If misclassifications are smaller in number then they can be ignored and also readjusting the margin can be avoided. In that case the margin gets re-designated as soft margin. Through the concept of cross validation better soft margin can be determined. Cross validation determines optimal number of misclassifications and observations allowed within good soft margin. This soft margin is called Support Vector Classifier. This name comes from the fact that this soft margin is actually obtained from edged data points with minimum overlap of two groups.

Minimum overlap is to be considered in the sense that few misclassifications allowing better classification in the long run. For 2-dimensional space line fit does the classification clearly. For 3-dimensional space plane fit does the work. For n-dimensional space hyper plane fit does the work. Clear classification of dataset is not possible due to tons of overlaps in some low dimension, and then transforms it into higher dimension. Cross validation concept is used to decide right fit of some higher dimension. If no nice linear classifier exists for clear classification data under consideration then Support Vector Machine (SVM) facilitates moving data to relatively higher dimension and finding relatively high dimensional support vector classifier to separate them into two categories effectively. However, transformation of data from relatively low dimension to relatively higher dimension involves huge computations. Further such a transformation is cumber some and complex. Here Kernel functions come handy.
There are several kernel functions. In this research article Radial Basis Kernel is chosen for the reason that it can transform some low dimension data into infinitely higher dimension without actually doing the transformation but facilitating in getting higher dimension parameters with simple calculations on relationships of data points. Thus RBF allows to find support vector classifier in infinite dimension space. Actually Radial Basis Kernel behaves like weighted nearest neighbor model. In other words, RBF considers the point that closest data points will have more influence on the classification of target data point under consideration rather than that of farthest data points of it.

From dataset, we focused mostly on parameters, such as, approx_cost_of_dining for a group of people, average item price, overall rating, and average monthly orders. These parameters obviously play major role in helping food startup to come up with right mix of restaurants and in turn serve potential customer in a better manner. As a sample, analysis of approx_cost_of_dining for a group by each classifier is presented in the next section of this article. Percentage of attainment of this parameter for different time periods is tabulated first. Later charts are drawn to facilitate user to visualize and easily understand these results.

VI. EXPLORATORY ANALYSIS

Using concept of cross validation dataset is divided into training and testing parts. Classifier of each algorithm mentioned in the previous section is designed and its learning and predicting accuracies are estimated.

Algorithm 1: The outcome of RBFSVM are tabulated and plotted as follows.

Table. I: Outcome of RBFSVM

| Time Period (days) | Training Set Length | Studying Precision | Forecast Precision |
|--------------------|---------------------|--------------------|--------------------|
| 10                 | 5000                | 80.90              | 70.25              |
| 40                 | 21000               | 81.51              | 70.45              |
| 80                 | 32000               | 82.31              | 70.85              |
| 150                | 51000               | 82.95              | 70.98              |

Charts depict accuracy of studying and predicting phase values in table 1 are given as below.

Fig.1. Chart depicting precision of study phase
From the close observation of these graphs it appears that over-fitting happens at smaller time period. However, it gets vanished to substantial extent with increase in time period. It means that larger the time period the more accurate and correct the fit will be. Over-fit causes small bias in learning and high variance in prediction results. This will lead to faulty class labels. This can be pruned using proper bias variance tradeoff.

Algorithm 2: The outcome of LSVM are tabulated and plotted as follows.

Table. II: Outcome of LSVM

| Time Period (days) | Training set length | Studying Precision | Forecast Precision |
|--------------------|---------------------|--------------------|--------------------|
| 10                 | 5000                | 86.32              | 64.35              |
| 40                 | 21000               | 86.36              | 65.80              |
| 80                 | 32000               | 87.95              | 66.90              |
Charts depict accuracy of studying and predicting phase values in table 1 are given as below.

Algorithm 3: The outcome of LR are tabulated and plotted as follows.
Table. III: Outcome of LR

| Time Period (days) | Training set length | Studying Precision | Forecast Precision |
|-------------------|---------------------|--------------------|-------------------|
| 10                | 5000                | 62.30              | 63.30             |
| 40                | 21000               | 62.88              | 64.10             |
| 80                | 32000               | 64.20              | 65.98             |
| 150               | 51000               | 65.30              | 66.95             |

Charts depict accuracy of studying and predicting phase values in table 1 are given as below.

Fig.5. Chart depicting precision of study phase

Fig.6. Chart depicting precision of forecasting phase
The observation of all these charts shows that accuracy percentage in learning and prediction is becoming better from algo 3 to algo 1. Further imply that precision percentage is far lower than benchmark considered in algo3 and algo2. In algo 1 the precision percentage is reaching the benchmark very decently. Note that precision percentages are increasing in studying and forecasting phases for all the three algorithms. In algorithm 3 (LR), two dimension space can be considered and best line is then fitted. In algorithm 2 (LSVM), two or three dimension space can be considered and best line or plane is then fitted. In algorithm 1 (RBFSVM), infinite dimension space can be considered and best hyper plane is then fitted. Algorithm 1 allows to have more parameters as well as combination of direct and indirect parameters for analysis and prediction [10]. Over-fitting gets reduced drastically in case of algorithm 1. It is visible as its results are converging to the benchmark very decently. These results can be further pruned by doing bias and variance tradeoff. The preferable factor is that bias should be a bit more in learning but variance should be far less in prediction. This factor is clearly met in case of algorithm 1. Thus SVMRBF is producing better results than the results of other two algorithms.

During this research work it is also noted that precision improves in both the phases if both direct and indirect parameters of dataset are taken in to consideration.

Table. IV: Forecast Precision in percentage for three algorithms are tabulated as below.

| Time Period (days) | Forecast Precision in % for algo 1 | Forecast Precision in % for algo 2 | Forecast Precision in % for algo 3 |
|--------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| 10                 | 70.25                             | 64.35                             | 63.30                             |
| 40                 | 70.45                             | 65.80                             | 64.10                             |
| 80                 | 70.85                             | 66.90                             | 65.98                             |
| 150                | 70.98                             | 67.90                             | 66.95                             |

The above table depicts that forecast precision in percentage is reaching the benchmark required in case of SVMRBF (algorithm 1). In other two algorithms it is not reaching that benchmark.

**VII. METHODOLOGY**

Procedure of designing of required system is given as below.

Step 1: Take the set of data with both basic and induced characteristics of food shop. Through concept of cross validation segment it into three parts, like too old, recent old, and present.

Step 2: Make Predictors of three algorithms one by one. See them get trained with corresponding data kept for the purpose. Observe training precision of the results.
Step 3: Check whether precision is up to the benchmark. Fine tune it if is not up to the mark by adjusting weights associated with the parameters. Once this requirement is met then go to next step.

Step 4: Then make these predictors to do prediction on test data. Check for required attainment of precision in prediction. If it is not reached then make to improve using bias and variance tradeoff.

Step 5: Pick the best predictor of the three. Develop proposed recommender system using that predictor model and finally deploy it to eatery outlet for use.

VIII. FLOWCHART

Consider direct and indirect characteristics, like, average_cost, item_price, rating, number of online orders per day and popularity.

Divide dataset into past, near past, and current parts

Train RBFSVM on past and near past data

Calculate RBFSVM Learning Accuracy

\[ \text{Accuracy} \geq \text{threshold} \]

Yes

No

Adjust weights and repeat

Proceed with testing phase of RBFSVM.

\[ \text{Accuracy} \]

Yes

No

Bias and Variance

Use this predictor model in developing the proposed recommender system.
IX. CONCLUSION

The following are finally inferred out of this research article.

1. RBFSVM predictor model is ideal choice for proposed system development.

2. Consideration of both direct and indirect parameters leads to better results than results of consideration of each of them separately.

3. Bigger the learning data smaller will be the over-fit. This factor definitely supports for obtaining better results.

X. FUTURE WORK

This work can be further extended in the following manner.

1. Try this work by considering descriptive nature of review dataset in place of binary nature.

2. Handling missing values of characteristics in different manner, instead of putting zero try to come up with guessed value nearest to it, will produce even more good results.

3. This work can be tried with other machine learning algorithms such as decision trees and so on.

REFERENCES

1. Zomato dataset online available: https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
2. C. M. M. de Figueiredo, "Emotions and Recommender Systems: a Social Network Approach", 2014.
3. Li, Y., Bontcheva, K., and Cunningham, H., “Adapting SVM for data sparseness and imbalance: A case study in information extraction. Natural Language Engineering”, 2009 15(2):241–271.
4. Ferwerda B, Graus M, Vall A, Tkal’ci˘c M, Schedl M, “The influence of users’ personality traits on satisfaction and attractiveness of diversified recommendation lists”. In Proceedings of the 4th workshop on emotions and personality in personalized services (EMPIRE 2016), Boston, USA.
5. Gunawardana A, Shani G, “Evaluating recommender systems”. In: Ricci F, Rokach L, Shapira B, Kantor PB (eds) Recommender systems handbook, chap. 8, 2nd edn. Springer, Heidelberg, pp 256–308, 2015.
6. Harambam, J., Helberger, N., & van Hoboken, J., “Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem”. Philosophical Transactions of the Royal Society A:Mathematical, Physical and Engineering Sciences, 376(2133), 20180088. https://doi.org/10.1098/rsta.2018.0088, 2018.
7. Friedman, A., Knijnenburg, B., Vanhecke, K., Martens, L., Berkovsky, S., & Berkovsky CSIRO, S. “Privacy Aspects of Recommender Systems”. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender Systems Handbook (2nd ed., pp. 649–688). New York: Springer Science + Business Media, 2015.
8. Marlin, B. M., Zemel, R. S., Roweis, S., & Slaney, M., “Collaborative Filtering and the Missing at Random Assumption”. UAI, 2007.
9. Xiao, J., Luo, M., Chen, J.-M. and Li, J.-J. ’An Item Based Collaborative Filtering System Combined with Genetic Algorithms Using Rating Behavior’, Journal of Advanced Intelligent Computing Theories and Applications, LNAI, Vol. 9227, 2015, pp. 453–46
10. Tommaso Di Noia, Jessica Rosati, Paolo Tomeo, and EugenioDiSciascio. “Adaptive multi-attribute diversity for recommender systems”, Information Sciences, 382:238:254–253, 2017
11. Meijun Liu, Xiao Hu, and Markus Schedl, “Artist Preferences and Cultural, Socio-Economic Distances Across Countries: a Big Data Perspective”. In Proceedings of the 18th International Conference on Music Information Retrieval, ISMIR 2017, 2017.
12. Friedman, A., Knijnenburg, B., Vanhecke, K., Martens, L., Berkovsky, S., & Berkovsky CSIRO, S. “Privacy Aspects of Recommender Systems”. In F. Ricci, L. Rokach, & B. Shapira (Eds.), Recommender Systems Handbook (2nd ed., pp. 649–688). New York: Springer Science + Business Media, 2015.
13. Dietmar Jannach and Lukas Lerche and Matthaus Gdaniec. Re-ranking Recommendations Based on Predicted Short-term Interests – A Protocol and First Experiment. ITWP 2013: Proceedings of the workshop Intelligent Techniques for Web Personalization and Recommender Systems at AAAI, 2013.
14. B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, “The new data and new challenges in multimedia research”. arXiv preprint arXiv:1503.01817, 2015.
15. The Recommender Systems: An Introduction by Dietmar Jannach, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich
16. Powers, D. M., Evaluation: from precision, re-call and F-measure to ROC, informedness, markedness& correlation.Journal of Machine Learning Technolo-gies, 2011, 2(1):37–63.
17. Paraschakis, D., “Algorithmic and Ethical Aspects of Recommender Systems in E-Commerce”, Malmö, 2018.
18. Xiao, J., Luo, M., Chen, J.-M. and Li, J.-J. (2015) ’An Item Based Collaborative Filtering System Combined with Genetic Algorithms Using Rating Behavior’, Journal of Advanced Intelligent Computing Theories and Applications, LNAI, Vol. 9227, pp. 453–46

AUTHORS PROFILE

Prof. (Dr.)Naga Raju M., Faculty in Computer Science Department, GITAM School of Sciences, GITAM (Deemed to be) University, Bengaluru - 562163, Karnataka, India. Domains of my research are Artificial Intelligence, Machine Learning, Deep Learning, Data Analytics and Cyber Security. I have got 8 publications till today. I have been felicitated by research awards consecutively for 2 years from 2016 by VIT, Vellore, Tamilnadu, India. I have been appreciated for my services by VIT, Vellore, Tamilnadu, India. I am member of CSI, IEEE.
