Predicting if a Customer will order food online again

Contributors: Shuvam Bose,Vaishali GaneshKumar,Vaishnavi N from PESU EC Campus, CSE.

Introduction & Background:

All over the world there's a sudden surge in the online purchase of consumables. Of which the food section leads. To interpret the same through the minds of the customers through their reviews and reactions pertained to the data set we have chosen. Hence we have decided to investigate as to which parameters influence the "Online Food Delivery Services" from the point of consumer perspective.

With this I would like to introduce you to our topic -"Predicting if a Customer will order food again".The data set has been taken from Kaggle and has a rating of 10/10. It has 55 attributes and 388 distinct records that help us in analyzing the same.

Previous Work/ literature Review:

Some references from our literature review for the analytics.

Paper1: This paper intends to predict and classify various reviews and produce their intuitions over them. For this they have manually collected 153 data by Non probabilistic, convenience data sampling method in different parts of an IT hub -Pune. Which is similar in aspects to our Bengaluru. They have mentioned about the model- "Technology Acceptance Model (TAM)" that they have used. To validate the same, they have used the infamous Hypothesis test (reliability test). They happen to reject their H0 as reliability coefficient >= 0.872. From which they draw the conclusion that -" Doorstep delivery" was the most appealing factor followed by "Ease & Convenience". They also drew the fact that Zomato outperformed its competitors in that region as per the survey results because of - "Rewarding, Location and Recommendations" features. Nothing was specified by them in their future scope.

Paper 2: This paper intends to perform sentiment analysis using DL. Their main aim is to predict the review of the customers. For different dataset heterogeneous models were used. DL techniques used were- LTSM and Bi-LTSM. Which was then verified using XAI -" technique against its computing logic". The accuracy of various models was tabulated in detail. Of which they concluded that DL models out-performed ML models for the same, especially for the Categorical data. Some DL models utilized were- CNN, XAI, ANN, Classification.

Paper 3: "Sentiment analysis on User reactions" was performed on the dataset using BERT Model. Their intention was to -predict the user reactions. Different data sets were used by their references' which weren't mentioned but were told could be provided on personal request. Some models used were- CNN, BERT, Graph-CNN, LSTM, SVM ,Bi-LSTM and many more. They have used a confusion matrix to validate their theory with overall model's accuracy- 92.86% which is a very significant figure as per the authors. Some additional details were well tabulated and concluded that the BERT model performed the best among other models that were used by them.

Paper 4: For estimating urban distribution of demand for short-term food delivery, a multi-target CNN-LSTM regressor has been developed. The research suggested assessing the distribution of demand for short-term food delivery across urban zones using deep neural network-based technologies. Using actual data from a food delivery business, the research focuses on hourly orders and frequent prediction updates. The objective of the sequential modeling strategy, which is based on a multi-target CNN-LSTM regressor trained on location-specific time series, is to detect sudden shifts and unexpected variations outside of the general demand trend. The methodology employs a single model for all service regions at once and a single one-step volume inference for each region at each time update. When compared to moving averages and other more traditional statistical techniques, the results demonstrate improved performance.

Paper 5: Using the Explainable Artificial Intelligence (XAI) Technique, "Unboxing Deep Learning Model of Food Delivery Service Reviews" - In this study, sentiment analysis was carried out in the FDS domain using simple and hybrid DL techniques (LSTM, Bi-LSTM, Bi-GRU-LSTM-CNN). Then, SHAP and Local Interpretable Model-Agnostic Explanations were used to explain the predictions. The Deep Learning models were trained and evaluated using the dataset of customer reviews that was extracted from the ProductReview website. The findings showed that the accuracy of the LSTM, Bi-LSTM, and Bi-GRU-LSTM-CNN models were 96.07%, 95.85%, and 96.33%, respectively. The model should show fewer false negatives because FDS firms work hard to identify and address every customer complaint. Over the other two DL models, Bi-LSTM, the LSTM model was chosen.

Paper 6: The model here is to classify and forecast the collected data from four Online food delivery companies from India and US. Regression models (OLS), LDA Tool and VADER based model are the Ml models used with Hypothesis test as the statistical method and text mining, regression analysis for validation. Important dimensions, a better service and customer satisfaction is inferred from the analysis. Comparisons of the different companies are studied during the pandemic. We can extend the scope to analyze OFD in different countries and include more financial variables. This approach provides high internal consistency and is acceptable as per validation approach.

Paper 7: Outcome of this paper is to predict customer purchasing decisions in the OFD industry. The dataset covers a variety of characteristics related to OFD providers which can be analyzed using predictive modeling with CART Decision tree, random forest and rule-based classifier measures across the cross-validation approach. This provides good accuracy of above 90% where single customers and males are more likely to use OFD services. Decision tree is easy to understand and fast but better to add enhancements to the nodes to increase output accuracy.

Paper 8: The outcome of this paper is to predict the cause of people to purchase through food delivery services and to analyze consumer value perspective. The data was collected from FDA consumers in the USA between the ages of 21 and 60 on which structural equation model is used with the hypothesis test and confirmatory factor analysis to validate it. All considered measures explained 29.59% of the variance which is lower than the threshold value of 50%. The results are the analysis of the consumer's decision-making process, associations of different variables and their impact. To improve the results, we can include comparative studies, add characteristics, try different sampling procedures, and use customer experience frameworks.

Proposed Solutions:

(a) **Pre-processing:** this involved data cleaning- of the reviews column which had null values. Binning of the Age attribute and conversion of categorical variables to numeric has been done. Some attributes such as 'Latitude', 'Longitude' and 'Pincode' have been dropped as they were not relevant to the problem statement.

(b) **Building Models:** The models used were Random Forest Classifiers for feature extraction. The prediction as to whether the customer would visit again (based on features extracted) was done using Decision Tree, Random Forest Classifier, Decision Tree with pruning and churn analysis.

(c) **Evaluation:** We have taken adequate precaution to ensure that we neither Overfit nor Underfit our models as the concern was raised by the Teaching Assistant. Hence we have decided to evaluate our models based on distinct training and test data sets. For that we have used random sampling with stratified sampling. We have also added adequate test statistics for each of our models to strongly justify our models accuracy. We have used parameters such as F1 score,Accuracy,confusion matrix and many more. The detailed test results have been added under the "models" section of the paper .Kindly refer to them. It has been done to prevent unwanted increase of pages and redundant data. We justify our models' accuracy with the High F1 score we have achieved for each model (>95 for random forest & ~90 for Decision tree).We have also added the confusion matrices for justifying our models that have adequate TruePositives and minimum FalsePositives. Hence we have an accuracy of 95+ % in each of our models.

Answers to Peer Review:

Q1) What attributes have been discarded from your dataset and why ?

A) We have discarded attributes such as - "latitude, longitude and pincode". Because we felt that it really doesn't contribute to our problem statement to determine which parameters affect the decision making procedure of our customers. Also we have taken into consideration only certain attributes that have maximum information gain (IG).

Q2) How did you come up with these models that you have presented?

A) We have come across multiple research papers that have analyzed a plethora of models. Out of them we choose to implement Random Forest and Decision trees because they happen to predict/perform better than others .

Q3) Why did you choose ten features when you have such little data? Only about 350 rows. Do you think your model will perform well with so little data and ten features?
(or)

Why did you choose the threshold of 0.025 ? Why not higher ? What influenced you to take such a decision?

A) As per our understanding we felt that training any model with 55 features and only 388 approx rows would result in overfitting. So we didn't have any fixed value for the threshold in our mind , we just wanted to train with only attributes that mattered our customers the most. Also our main motive was to prevent overfitting the model to our data set.

Recommendations given by the teaching assistant:
i) Use fewer features about 3 or 4 but not more than that as it looked overfitting to him.
ii) Print the model report to better understand the condition of the model(if it still overfits or not).
iii) Show the F1 score, recall, precision.
iv) Add more evaluation metrics to have more information about the model.

Explanation of each Component:

Pre-processing: The Reviews column was found to have missing information (such as 'Nil', 'No comment') this was replaced with NaN value for uniformity. Then the dataset was checked for Null values- only the Reviews column was found to have 147 Null values. These Null values have been replaced with 'Not specified'. Binning has been done for the age attribute. The relevant categorical attributes have been converted to numeric type. Unwanted columns such as Latitude, Longitude, Pincode have been dropped.

EDA: *Outlier analysis:* Age was the only attribute that had a few outliers. This did not pose a problem as the Age attribute has not been taken into account in the model.
Demographic analysis: Men have ordered more than females, Customers of the age of 23 have ordered the most, followed by age 22, Most of the customers are students
•Majority of the customers are Post graduates, 48.2% of the customers have no monthly income.
Customer Preferences Analysis: Customers prefer food delivery apps over walk-in, phone call and web browser, Most customers prefer to order snacks, Non-veg was found to have more preference, No specific order time (weekend/weekday) was observed.
Factors influencing next order: Most of the customers value rating of the restaurant, Freshness and taste of the food are found to be very important, Temperature is also an important factor to majority of the customers.
Correlation Plot: Output was found to have significant positive correlation with 'Ease and convenience', 'Time saving', 'More restaurant choices' and negative correlation with 'Self Cooking', 'Health Concern', 'Late Delivery'.

Models: Random Forest classifier was used for feature selection based on importance. The top 4 important features were selected for prediction models.
The models used for prediction:
Random Forest classifier, Decision tree, Decision tree with pruning and churn analysis.

Experimental results & Explanations:

Models:

Random Forest Classifier:

```
Test Results:
precision    recall  f1-score   support

         0       0.93      0.96      0.94       116
         1       0.86      0.78      0.82        40

 accuracy                           0.91       156
 macro avg     0.89       0.87      0.88       156
weighted avg  0.91       0.91      0.91       156

Confusion matrix:
[[111    5]
 [  9    31]]

Accuracy:
0.9102564102564102
Precision:
0.9250000000000000
Recall:
0.7816901408450704
```

Decision Tree:

```
Test results:
precision          recall  f1-score   support
         0         0.96      0.89      0.93       325
         1         0.60      0.83      0.69        63
accuracy                              0.88       388
 macro avg       0.78      0.86      0.81       388
weighted avg    0.90      0.88      0.89       388

Confusion matrix:
[[290              35]
 [ 11              52]]

Accuracy:
88.14432989690721
Precision:
0.963455149501661
Recall:
0.847953216374269
```

In general wrt models we have concluded that the models perform best when the data are normalized and have more training data to make the models understand and learn the data. The models gave overfitting results when more number of attributes(> 4) are taken into account.

We have pruned the trees to prevent overfitting . And to prevent bias we have used random stratified sampling where ever necessary.

Conclusions:

From the above we conclude that our interpretation of the dataset aligns with the Research papers that we have referred to to get a direction to follow.

The EDA gives us insights on the demographic details to the most frequent consumers.

We come to know that there is no statistically significant (>= 0.5) correlation among the attributes from the correlation chart. Most orders are placed by men rather than women. The youth (age range of 22-23) order the most and most of the consumers possess Post graduation level of education. Hence 48.20% of the respondents have no monthly income.

Also factors like Rating of the restaurant, taste & quality of food and the temperature of the food matter to our customers the most.

And we also concluded that consumers prefer snacks to be delivered the most, non-veg was ordered the most and no specific order time was noticed.

From the models we concluded that the “Random Forest Classifier” performs better than “Decision Tree” by a few units.

Contribution of each member:

Predicting if a Customer will order food online again

Contributors: Shuvam Bose,Vaishali GaneshKumar,Vaishnavi N from PESU EC Campus, CSE.

Shuvam Bose (PES2UG20CS[redacted]): Building and fine tuning the models. Perform the basic EDA. Referred to three research papers for literature review.

Vaishali GaneshKumar(PES2UG20CS[redacted]): Referred to two research papers, worked on preprocessing, EDA and basic model building.

Vaishnavi N(PES2UG20CS[redacted]): Referred three research papers, EDA and basic model building .

References:

Literature review references:

Paper-1	https://5y1.org/download/3a0882d005f3184157c3fe6d19f8b170.pdf
Paper-2	https://www.mdpi.com/2304-8158/11/10/1500
Paper-3	https://ieeexplore.ieee.org/document/9441669/citations#citations
Paper-4	https://www.mdpi.com/2304-8158/11/14/2019/htm
Paper-5	https://www.sciencedirect.com/science/article/pii/S014829632200159X#s0010
Paper-6	https://www.sciencedirect.com/science/article/pii/S096969892200145X
Paper-7	https://arxiv.org/abs/2110.00502
Paper-8	https://www.sciencedirect.com/science/article/pii/S0969698921002332