The Prediction of Feedback with Demographics & Locations of Users in Online Food Delivering Based on Machine Learning Models

Junyi Hu
Institute of Science of Mathematics, Nanjing Normal University, Nanjing, China

Keywords: Prediction, Demographics & Locations, Online Food Delivering, Machine Learning Models.

Abstract: The online food delivery (OFD) becomes a hit in the previous decade. However, the OFD companies face different challenges in reality. With the data providing demographic features and locations, this research makes use of data learning models to predict the feedback of the OFD users, which implies the probability of becoming a long-term user. This research takes ANN neural network, Decision Tree, Random Forest into consideration, and make comparison of the performance of them, and all the models reach the accuracy above 80%. The geographic feature of the users is considered in an abstract map, and the geographic center of the user with negative feedback is to the north-west of the geographic center of all the users. In addition, feature analysis reveals the age of user is the feature with most importance, followed by family size. The performance of the models is estimated by different evaluating indicators, such as confusion matrix, f-measure and accuracy.

1 INTRODUCTION

With the advancement of information technology, the widespread application of OFD services is evident. However, not everyone readily embraces this new trend. Therefore, investigating the primary user demographics of OFD services is a worthwhile research endeavor. Factors such as Information Quality (IQ) and Promotions (PRO) significantly influence whether an individual becomes an online food delivery user and their satisfaction with the service. Accurately predicting the likelihood of someone becoming a potential user of online food delivery services based on individual factors is crucial (Moroz and Polkowski, 2016; Prasetyo and Tanto, 2021 & Tan and Kim, 2021). Additionally, identifying which features are the most influential factors is of utmost importance (Shukla and Deshpande, 2023 & Wang et al., 2018). It can help the retailers and OFD platforms figure out the most likely long-term users and make better programming and advertising.

Previous research on OFD services has primarily focused on macro-level aspects, such as consumer attitudes towards these services and the structural components of the OFD system, including factors like convenience motivation and the perceived usefulness after usage (Yeo et al., 2017). However, these studies often overlook the nuanced details, particularly those concerning individual users within the OFD service ecosystem. They tend to treat OFD users as a monolithic entity, disregarding the intricate internal structure of consumer communities, which is a critical aspect addressed in this current research.

The neglect of individual user experiences and behaviors within the broader OFD consumer base can lead to a lack of understanding of the diverse motivations, preferences, and challenges faced by different user segments. For instance, some users might prioritize speed of delivery, while others might be more concerned with the quality of food or the sustainability of the delivery process. Similarly, the frequency of OFD usage, the types of meals ordered, and the time of day when orders are placed can vary significantly among users, reflecting diverse lifestyles and needs.

Moreover, the socio-demographic characteristics of users, such as age, income level, and geographical location, can influence their interactions with OFD services. Younger, more tech-savvy consumers might be more inclined to use mobile apps for ordering, while older users might prefer traditional methods like phone calls. Urban dwellers might have access to a wider variety of restaurants and faster delivery times compared to those in rural areas.

Hu, J.
The Prediction of Feedback with Demographics & Locations of Users in Online Food Delivering Based on Machine Learning Models.
DOI: 10.5220/0012982500004601

In *Proceedings of the 1st International Conference on Innovations in Applied Mathematics, Physics and Astronomy (IAMPA 2024)*, pages 13-21
ISBN: 978-989-758-722-1

This paper investigates the features that make an individual a potential consumer of OFD services. By employing supervised machine learning models and incorporating feedback from OFD users, we aim to delineate the profile of potential OFD users. Recognizing the significant impact of dataset quality on accuracy based on previous research, this study utilizes a dataset that includes geographical locations and demographic information related to online food delivery. Extensive efforts were made to portray a general image of OFD service users using this data. Multiple models were implemented in this study, and a comprehensive comparison of these models was conducted to leverage their strengths and address their weaknesses. Additionally, we placed particular emphasis on geographical factors such as longitude and latitude, as well as demographic factors like household size, education level, and monthly income. To present the findings clearly and intuitively, machine learning models and data visualization techniques were emphasized.

The dataset comprises rich independent features such as latitude and longitude, gender, occupation, household size, and individual feedback. The aim of this research is to analyze relevant data through machine learning models to comprehensively depict the profile of potential OFD users, thereby uncovering common characteristics and needs within these user groups. This approach not only provides guidance for OFD platforms to customize their services effectively and improve customer satisfaction but also offers insights for refining marketing strategies and user interface designs. By thoroughly analyzing these factors, we hope to enhance the understanding of diverse user group needs and thus support the development of the online food delivery industry.

The remainder part of this paper is constructed below. The second section is about the related work about the OFD service. Section 3 is mainly about the research method and detail of the dataset. The machine learning model applied is discussed in Section 4. Section 5 is about the result and the discussion of this experiment. Finally, we discuss the conclusion of the result and future work about OFD services.

2 RELATED WORKS

The satisfaction of the user of online food delivery system is determined by different factors. Previous studies on factors affecting satisfaction with online food delivery services have spanned a wide range of research areas. Machine learning witnessed a vast progress in the last decades, applied in extensive aspects of researches. Common machine learning methods include random forest, XGBoost, decision trees, artificial neural networks, Bayesian networks, and so on. (Yeo et al., 2017) Some studies trained their models on datasets and employed various machine learning models for prediction. Their research achieved high accuracy. Huycock Tan (2021) initially implemented a linear regression model, providing insights into online food delivery users. Their model focused on online food delivery services during the COVID-19 pandemic, illustrating factors that positively impact OFD user satisfaction. Another study by Janmejay (2024) applied a decision tree classification model, uncovering preferences of OFD users, offering better decision guidance for online food delivery aggregators. SVM and Ridge regression was applied in the research of Wang, W.M. (2018).

So far, the world of auto machine learning models and deep learning models has enlarged greatly. For example, PLS-ANN (Foo et al., 2018) and Bi-LSTM (Tam and Tanriöver, 2021). They have been successfully applicated into a wide range of fields and gain outstanding results.

3 METHOD

In this research, dataset is preprocessed into the form which the machine learning models and neural network are easy to manipulate on. machine learning techniques are implemented to predict the likelihood of potential user based on the features given. Optimization are conduct to improve the performance of the models, especially for the neural network. Based on the algorithm of random forest, the influence of different features on the result are conducted (Figure 1).

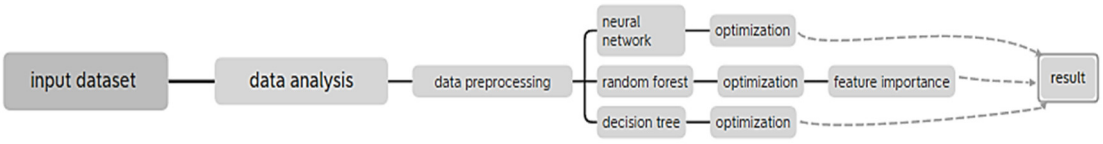

Figure 1: Research workflow (Picture credit: Original).

Table 1: Data Info.

	Age	Gender	Marital Status	Occupation	Monthly Income	Educational Qualifications	Family size	latitude	longitude	Pin code	Output	Feedback
0	20	Female	Single	Student	No Income	Post Graduate	4	12.9766	77.5993	560001	Yes	Positive
1	24	Female	Single	Student	Below Rs.10000	Graduate	3	12.977	77.5773	560009	Yes	Positive
2	22	Male	Single	Student	Below Rs.10000	Post Graduate	3	12.9551	77.6593	560017	Yes	Negative
...	...	...	...	...	...	...	...	...	...	...	...	...
387	23	Male	Single	Student	No Income	Post Graduate	5	12.8988	77.5764	560078	Yes	Positive

3.1 Data Analysis

Before the data preprocessing, having a rough understanding about the dataset by data analysis is crucial. The features and the data type are shown in table 1.

3.2 Data Preprocessing

Before building and training the data in the model, data preprocessing is necessary. To transform the data into the form which is easily implemented on, the labels of features, such as occupation, marital status are turned into categories labelled with number.

Focusing on the importance of different features, we need to transform the scale of different features with the standard score. It dismisses the value of each figure. Instead, the process of the z-scoring focuses on the distance of every point and the average of the feature. It follows the formulas given.

$$z=\frac{x-\mu}{\sigma} \quad (\mu=\frac{\sum_{i=1}^{n} x_i}{n}, \sigma=\sqrt{\sum_{i=1}^{n}(x_i-u)^2}) \quad (1)$$

3.3 Location Analysis

The feedback distribution by location is shown in the figure 2.

The location of the users is the central part of a small town in India. Firstly, the data show that the users of OFD are concentrate around the central point with Latitude 12.97 ° N and Longitute77.60 ° E. Density of the users decrease with distance away from the central point. In the plot, we can figure out the majority of the users with negative feedback lie in the area whose longitude in the range of 77.55° E and 77.62° E Latitude between 12.92° N and 13.07° N, and the central point of users with negative feedback is (12.99° N,77.57° E). In addition, the central point of users with positive feedback is (12.96° N ,77.64° E), to the south-east of the central point of negative-feedback-uses.

4 MODEL DESIGN

4.1 Neural Network

A neural network is a model imitating the structure and effect of the human brain. It consists of interconnected vertices, called neurons(nodes), organized into layers (Agatonovic-Kustrin and Beresford, 2000). Each layer is composed by multiple artificial neurons or processing elements, connected with weights(coefficients). A neural network typically consists of input layer, hidden layer and output layer.

It is easy to see the input layer takes in the data, and the output layer gives the result. The activate function, typically non-linear is applied to deal with the data given by the neurons in the previous layer, and transform the data into different signals, which is not seen by the outside. Such process is implemented in the hidden layers. The model training happens in the form of adjusting the weight of the connection between different neurons and layers (Figure 3).

In this research, the ANN model takes for layers of neurons, with the input layer with 3 neurons and the output layer with 2 neurons. The hidden layer contains 16 neurons each. Every layers are connected by ReLU activation function.

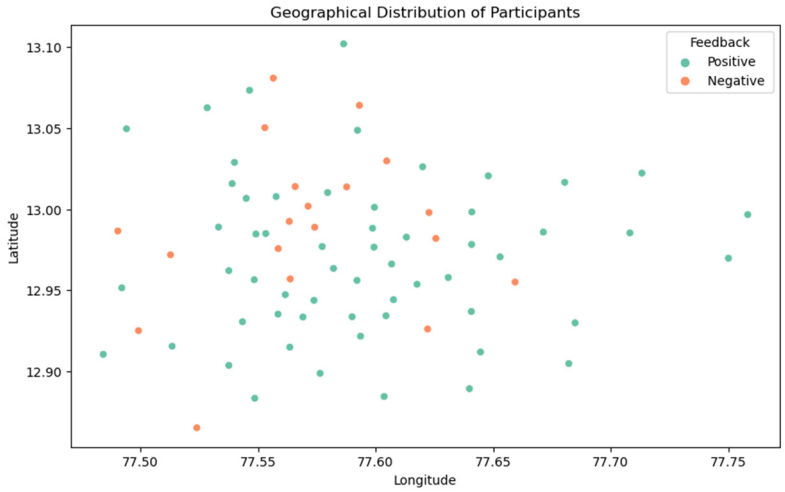

Figure 2: Location (Picture credit: Original).

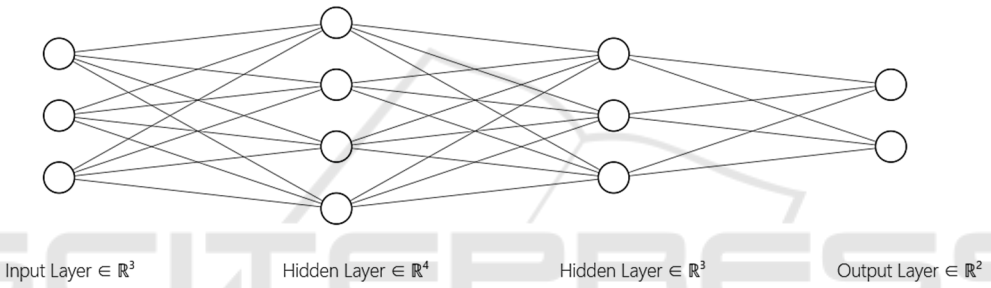

Figure 3: Neural Network (Agatonovic-Kustrin and Beresford, 2000).

In this research, there are three fully connected layers in the ANN neural network. The ReLU activation function is applied between different layers and applying drop out after each activation function. BCEwithLogistsLOSS play the role of loss function, which combines a sigmoid activation function and the binary cross entropy loss. The optimizer is ADAM with learning rate 0.0001.

In ANN the optimization is about the activation function. The sigmoid function, arctan function and ReLU function were used. Because of the similarity of the sigmoid function and arctan activation function, the results are very similar. The ReLU activation function perform best.

4.2 Decision Tree

Decision tree is a basic model in machine learning, based on classification and regression. In the decision tree there is an important value that describes the degree of internal difference and uncertainty of a category, which is typically called entropy. The smaller the entropy, the better the effect of classification. The value of entropy is given by the following formula

$$\mathbf{H(x)}=-\sum_{i=1}^{n} p_i * \log(p_i) \quad (2)$$

The decision tree model divides the dataset into several parts by one of the features. Then the model computes the entropy and the information gain ratio of different categories. The feature with the highest information gain ratio is applied first. With the first feature, the model can make the first decision, which classify the dataset into several categories with the lowest entropy. Now the first layer of the tree is finished, but the classification is not precise enough to make the prediction. In order to improve the accuracy, the process above is repeated. The model finds the best feature apart from the feature applied above. And the second layer is formed. However, if all the feature were used, the model may have the problem of overfitting. Consequently, some optimization such as pruning and make limitation on the number of layers of the tree.

In this research, the criterion is chosen as the entropy. The max depth is decided to be 10.

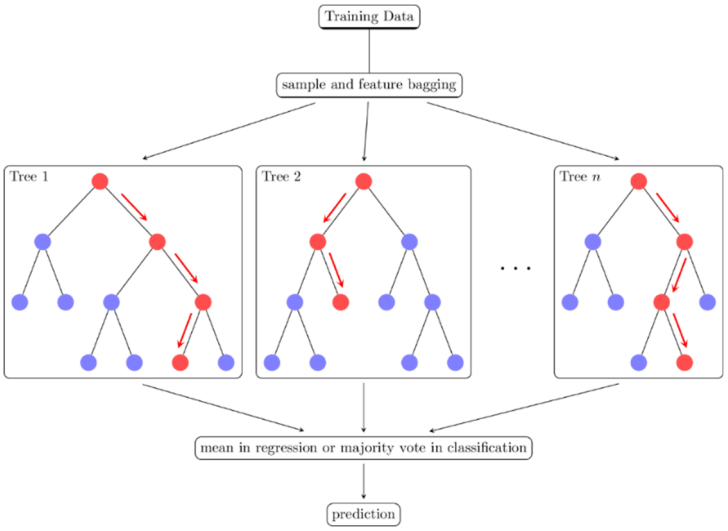

Figure 4: Random Forest (Wang et al., 2018).

At first the decision tree model performed the worst, for both the feature of the dataset and the model overfitting. In order to tackle with the problem, we first applied different algorithms including CART, C4.5 and ID3. Then we limited the depth in order to prevent overfitting.

4.3 Random Forest

Random forest is a state-of-art ensemble machine learning model. The essence of random forest is generating multiple decision trees (Figure 4). The model picks up a subset of k-features of the features give, typically k=log(n), and n is the total number of the features. The importance of each feature can be computed in the process of random forest predicting, which is a great advantage of the model. There are many classification trees in a random forest. We need to classify an input sample by inputting it into each tree for classification. Finally, the mean or majority vote in classification is the result.

In this research, the random forest is provided with a set of parameters and is allowed to choose the parameter to gain the highest accuracy. The result of the parameter decision is the n-estimator is 1000 and the max depth is 10.

5 RESULT AND DISCUSSION

This research put eyes on the performance of different machine learning models on the dataset to make prediction on the feedback of the user according to the demographic features and locations. This research take accuracy, confusion matrix, recall F-measure as the evaluation criterion. Accuracy is the most important feature about a prediction model, while recall ratio and confusion matrix helps the optimization of the models.

In this research, all models were implemented on python 3.10.9 environment, with Numpy, Pandas, Scikit-Learn and Pytorch packages. The hardware configurations comprise a 3.20 GHz AMD7 5800H CPU, a RTX 3060 GPU.

5.1 Dataset Analysis

The paper utilizes the online food dataset from Kaggle. Each input data consists 12 attributes. We explore the correlation of between the attributes completely, in order to find out the feature that are conducive to the predicting of feedback. A heat map was given in figure 5.

In this figure 6, we can conclude that most users of online food service gave positive feedbacks. The age distribution is the young people with age lower than 25 is in the most favor of the service. Especially the 29-year-old and 30-year-old population give more negative feedback than the positive, which is rare in this research. Male users take a higher proportion than the female, and they contribute a higher satisfaction rate. The majority of the marital status of the users is single, and the students show most interest in the online food service, particularly post graduate students.

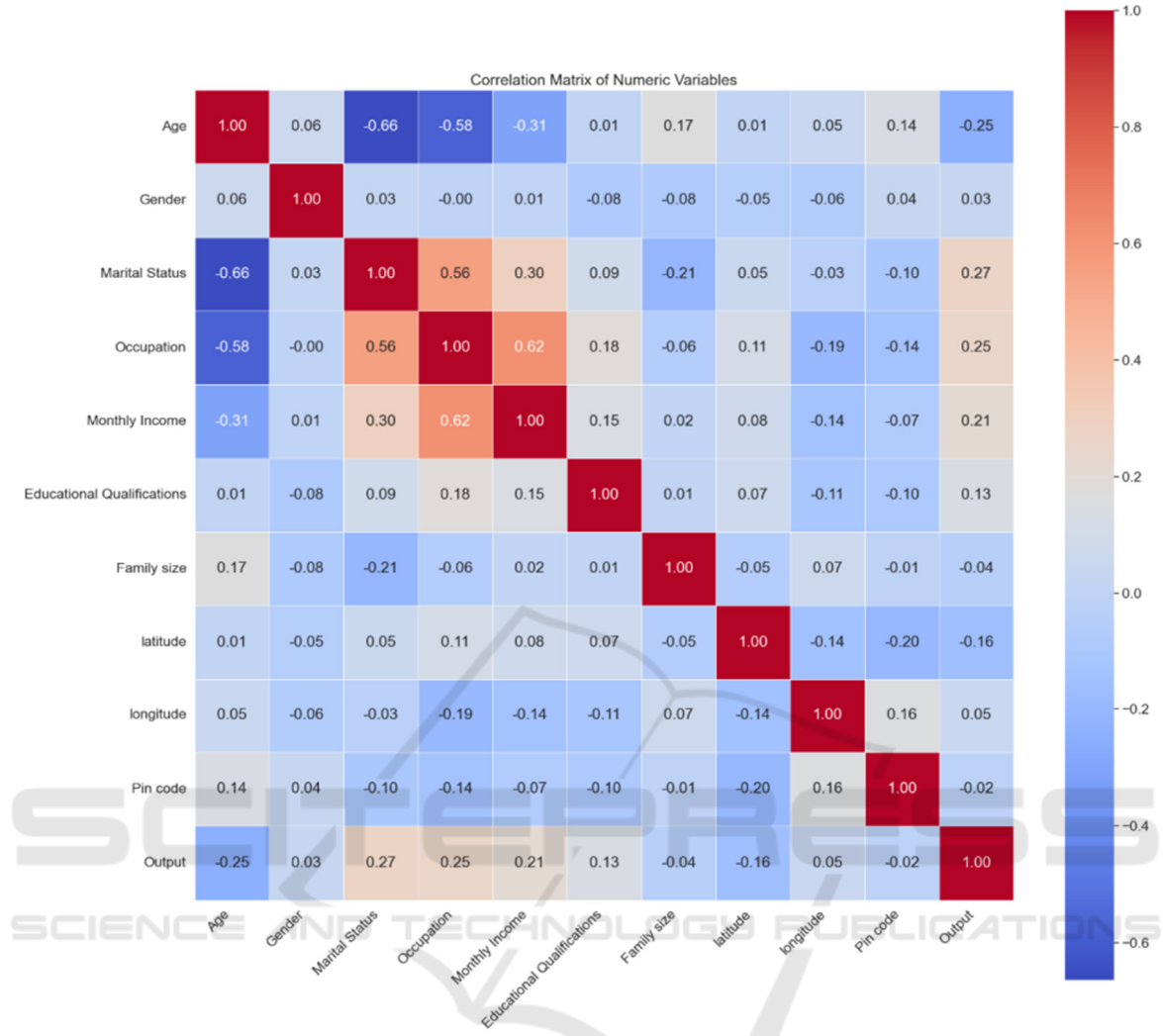

Figure 5: Heatmap (Picture credit: Original).

5.2 Model Prediction and Optimization

5.2.1 Confusion Matrix

The confusion matrix of machine learning predicting model is a crucial criterion of the performance of the 0-1 prediction models. It conveys the information about the actual data and the prediction intuitively.

The confusion matrices of the models are shown in figure 7.

5.2.2 Evaluation of the Models

The evaluation of the criterion on the models above are revealed in the table 2.

In the table we can get the conclusion that the accuracy of the three models are all above 80%, and the decision tree get the highest, while ANN is the lowest. The possible reason of the unsatisfying performance of ANN is the scale of dataset is too small. However, in the field of recall, ANN take the lead, with 90.47% came to the first. Decision tree also have the highest precision rate. The number of features of the data is in the performance advantage zone of Decision Tree model, which is a reason of the performance of Decision Tree. The differences of F-measure are at a small extent.

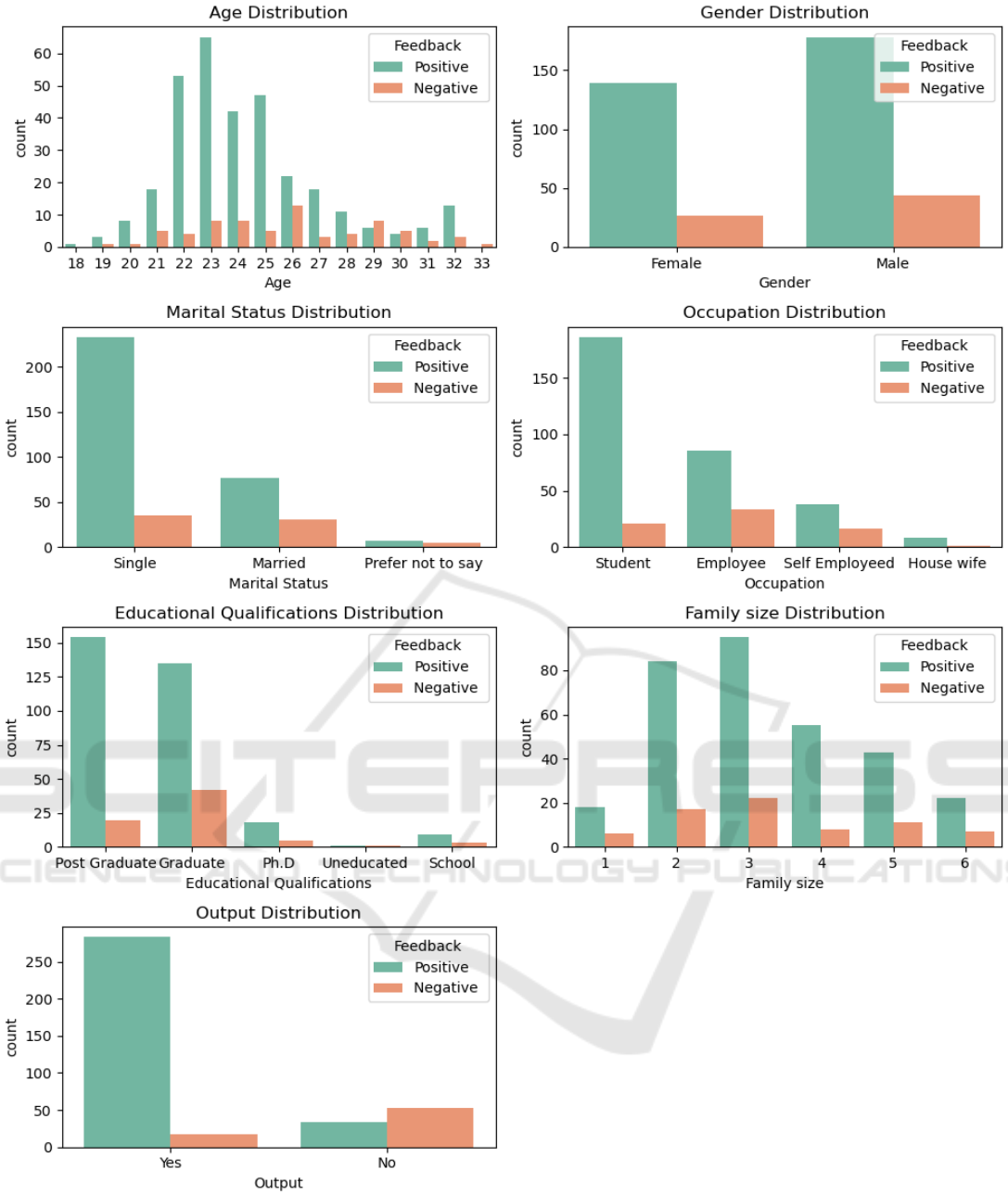

Figure 6: Counting of data (Picture credit: Original).

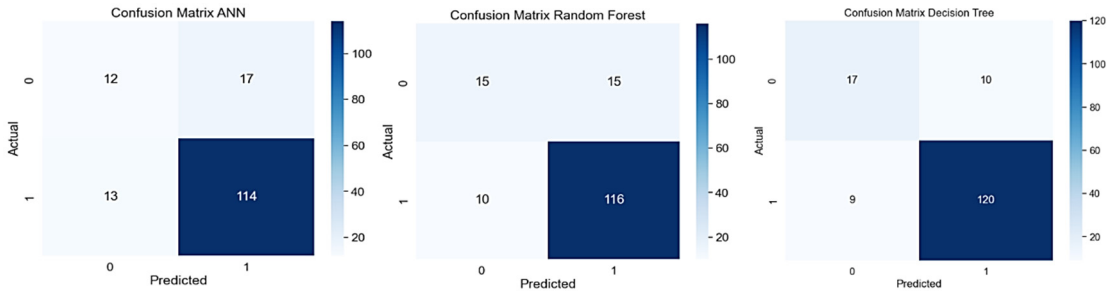

Figure 7: Confusion matrices (Picture credit: Original).

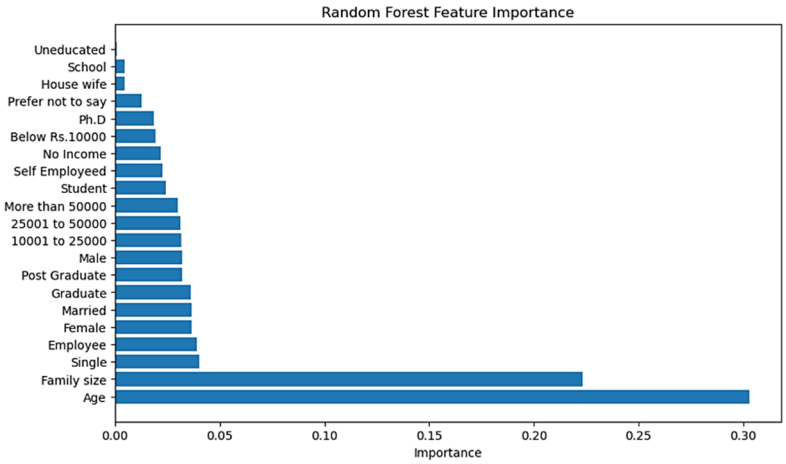

Figure 8: feature importance (Picture credit: Original).

Table 2: Models Evaluation.

Model	Accuracy	Recall	Precision
ANN	80.76%	90.47%	87.02%
Random Forest	83.97%	88.54%	88.54%
Decision Tree	87.82%	87.59%	92.30%

5.3 Importance Analysis

The random forest model has a great advantage. The model brings up multiple decision trees and take the average of the results of different trees. In this process, the distribution of each feature on the final result is shown clearly.

We divide the features into two classes. For the features with plenty of value, such as age, family size, are considered in category. However, for some features such as occupation are considered in the form of concrete feature. The result is shown in the figure 8.

In the see the age is the feature with highest importance, followed by family size. The importance of other features are far lower.

6 CONCLUSION AND FUTURE WORKS

In order to address the issue faced by OFD companies when determining potential users, this research utilizes a Kaggle dataset and machine learning models to predict user feedback based on demographic characteristics and geographic location. The study identifies age as the decisive factor in user feedback, providing guidance to companies on targeting specific audiences. All models achieved an accuracy of over 80%, with high overall evaluation metrics, indicating outstanding research results. Among the models used, Decision Tree outperformed others in various indicators, while the ANN performed moderately. This suggests that Decision Tree excels in binary prediction tasks, while ANN's performance is limited due to its two-node output layer. However, the research has certain limitations, including a relatively small dataset and homogenization bias towards users with positive feedback. Future research should consider larger and more diverse datasets, as well as explore optimization strategies for models, particularly for ANN models.

This research also has some limitations. The data is only about the OFD users in a town in India, which may lead to the deprivation of universality of the research. Future studies can widen the scope of dataset and trying to find the uniform regularity of OFD users. Additionally, the respective number of models applied is limited. Future researches can do some more optimizations and make use of a wider range of machine learning models to gain a more precise prediction.

REFERENCES

M. Moroz and Z. Polkowski, "The Last Mile Issue and Urban Logistics: Choosing Parcel Machines in the Context of the Ecological Attitudes of the Y Generation Consumers Purchasing Online", Transp. Res. Procedia, vol. 16, pp. 378-393, 2016.

Prasetyo, Y.T.; Tanto, H; Factors Affecting Customer Satisfaction and Loyalty in Online Food Delivery Service during the COVID-19 Pandemic: Its Relation

with Open Innovation. J. Open Innov. Technol. Mark. Complex. 2021, 7, 76.

Tan, H., & Eng Kim, V. W. (2021). Examining the Factors that Influence Consumer Satisfaction with Online Food Delivery in Klang Valley, Malaysia. The Journal of Management Theory and Practice (JMTP), 2(2), 88-95.

Shukla, J., & Deshpande, A. (2023). A decision tree classifier approach for predicting customer's inclination toward use of online food delivery services. Multidisciplinary Science Journal, 6(5), 2024072 https://doi.org/10.31893/multiscience.2024072

Wang, W.M.; Wang, J.W.; Barenji, A.V.; Li, Zhi; Tsui, Eric (2018). Modeling of individual customer delivery satisfaction: an AutoML and multi-agent system approach. Industrial Management & Data Systems, IMDS-07-2018-0279

Vincent Cheow Sern Yeo, See-Kwong Goh, Sajad Rezaei, Consumer experiences, attitude and behavioral intention toward online food delivery (OFD) services, Journal of Retailing and Consumer Services, Volume 35, 2017, Pages 150-162, ISSN 0969-6989,

Pik-Yin Foo, Voon-Hsien Lee, Garry Wei-Han Tan, Keng-Boon Ooi, A gateway to realising sustainability performance via green supply chain management practices: A PLS–ANN approach, Expert Systems with Applications, Volume 107,2018, Pages 1-14,

S. Tam, R. B. Said and Ö. Ö. Tanriöver, "A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification," in IEEE Access, vol. 9, pp. 41283-41293, 2021, doi: 10.1109/ACCESS.2021.3064830.

Agatonovic-Kustrin S, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. Journal of pharmaceutical and biomedical analysis, 2000, 22(5): 717-727.