STA 141 Final Project

Statistical Analysis and Predictive Modeling of Online Food Ordering Platform Feedback

Group Member Contribution:

Minyi Liu (Leader, myliu@ucdavis.edu): Part I, III, IV, V, Coding

Kieran Sullivan (khsullivan@ucdavis.edu): Part IV, V, VII, Coding

Manraj Dhillon (mmdhillon@ucdavis.edu): Part II, VI

Steven Lee (stelee@ucdavis.edu): Part III

Group 22

I. Introduction

In today's technologically advanced era, many innovations have brought convenience to our lives, and online food ordering platforms are among them. With just a gentle swipe on our smartphone screens, restaurants joining online food ordering platforms allow us to browse and order our desired foods from anywhere. It benefits those who are too busy with work or study to cook for themselves and helps those who need to be more skilled in cooking or simply prefer to sample the cuisine of a particular restaurant without going out. Regardless of the reasons, online food ordering platforms are a boon to humanity.

Our project utilizes data from the “Online Food Order” dataset obtained from Kaggle, aiming to explore and identify the most suitable predictive models for accurately predicting whether users of an online food ordering platform will use the service again based on their feedback and demographic attributes. This analysis will leverage a comprehensive dataset containing demographic, geographic, and feedback data to build and evaluate various predictive models. The dataset provides a comprehensive array of demographic attributes, including gender, income levels, educational attainment, family size, geographical data, and user feedback. These variables form the foundation for our analysis, providing a nuanced understanding of the factors influencing user behavior on online food ordering platforms. This, in turn, helps us understand which groups in contemporary society are more reliant on the convenience offered by online food ordering platforms.

Our objectives are twofold: first, we will use logistic regression to model the relationship between binary user feedback and other predictors, allowing us to forecast positive or negative feedback regarding the online food ordering experience. Second, we will use Linear Discriminant Analysis (LDA) to identify distinct user segments based on their preferences and behaviors. By comparing the performance of these two models, we aim to determine which approach better captures the dynamics of user preferences and provides deeper insights into user needs.

II. Research Problems

The data set gives us insight into several factors that might influence a customer's tendency to give positive or negative feedback to their overall food ordering experience. There are several factors at hand that can be considered. Through this project, we aim to find the following questions:

1) What demographic, geographic, and feedback attributes are significant predictors of whether users will use the online food ordering service again?
2) Which variables do not impact a customer’s rating tendencies to a significant extent?

3) Which model, logistic regression or LDA, provides more accurate and actionable insights into user preferences and the factors influencing their continued use of online food ordering platforms?

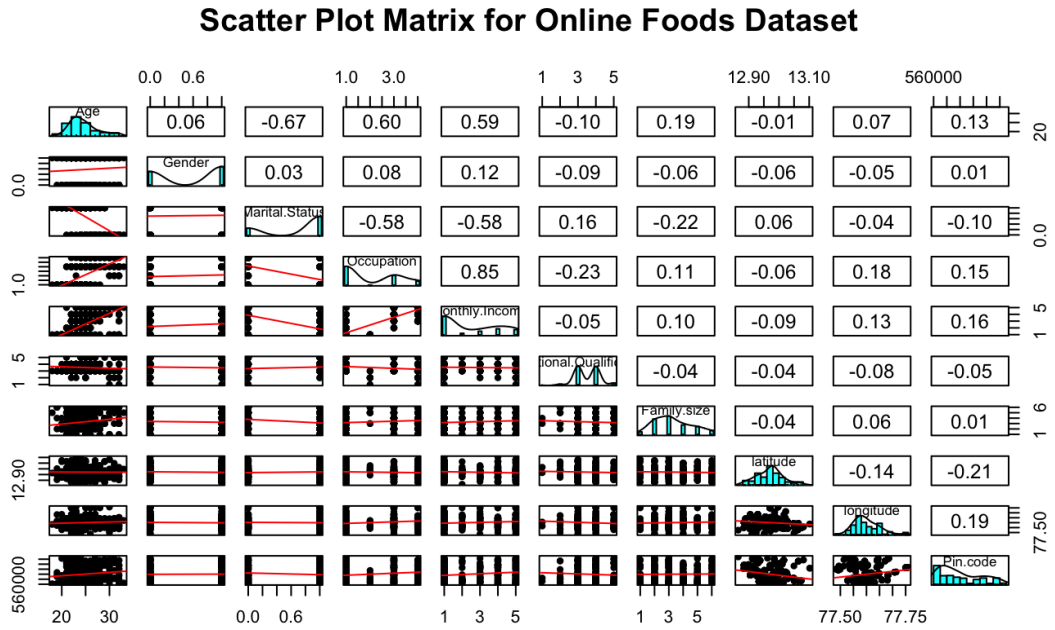
III. Data Information

The dataset for this project is obtained from an online food ordering platform and contains the following information:

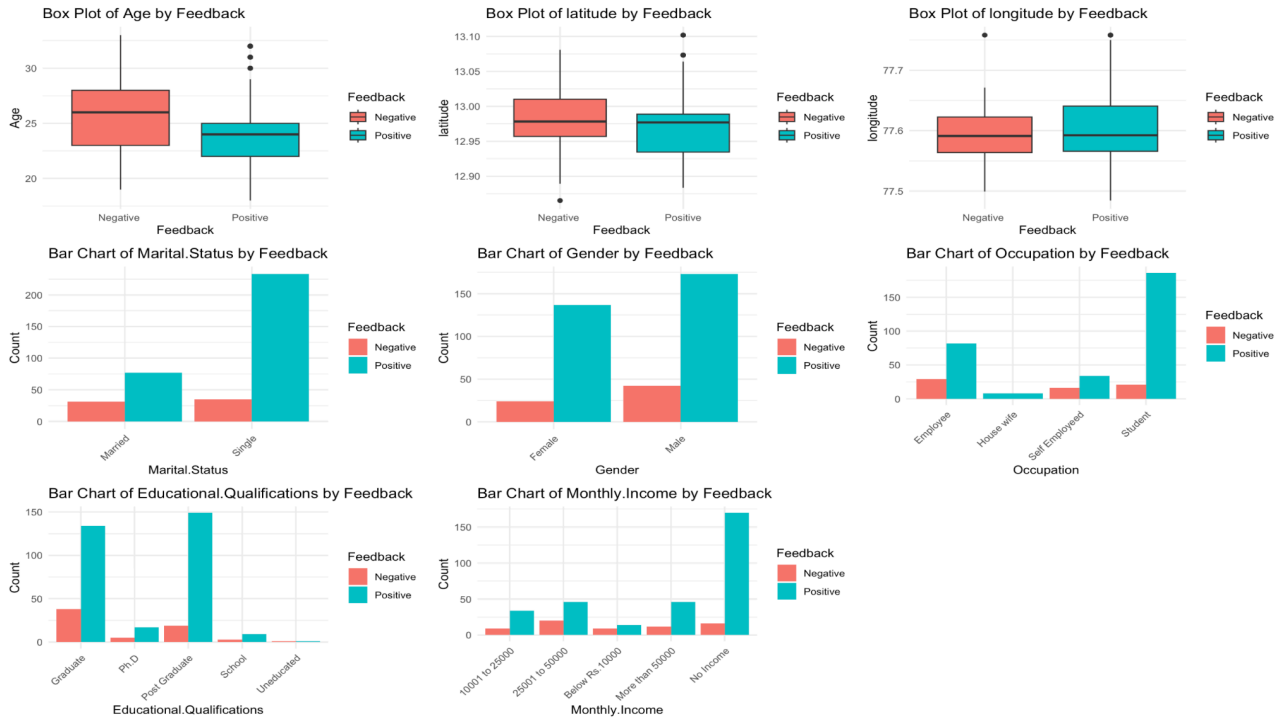
- Age: Numerical variable representing the age of the user.
- Gender: Categorical variable indicating the user's gender.
- Marital.Status: Categorical variable representing whether the user is single, married, etc.
- Occupation: Categorical variable indicating the user's job.
- Monthly.Income: Numerical variable representing the user's income per month.
- Educational.Qualification: Categorical variable representing the user's highest level of education.
- Family.size: Numerical variable indicating the number of people in the user's household.
- latitude: Numerical variable representing the user's geographic latitude.
- longitude: Numerical variable representing the user's geographic longitude.
- Pin.code: Categorical variable indicating the user's postal code.
- Feedback: Categorical variable indicating whether the user had a positive or negative experience with the platform.

During our preliminary data processing stage, our primary objective is to ensure the integrity and reliability of the dataset. We adopt a meticulous approach to address missing values, opting for their removal to maintain the dataset's quality. Categorical variables are transformed into numerical representations using the "as.factor" method, facilitating streamlined computation and analysis. Furthermore, in our efforts to refine the dataset for relevance to our project objectives, we have chosen to exclude the variable "Pin.code." After careful consideration, we have determined that this variable does not significantly contribute to the insights we seek to derive from our analyses. Moreover, to maintain data consistency and accuracy, we have decided to disregard any rows where the marital status is listed as "prefer not to say." By excluding such entries, we ensure that our analyses are based on reliable and representative data. As a result of these rigorous cleaning and refinement processes, our dataset now consists of 376 observations, each encompassing 10 variables.

IV. Data Visualization



This scatterplot matrix provides us with useful information on the distributions and relationships between our 10 original predictor variables. Running from the top left corner to the bottom right corner are histograms showing the distribution of the data for that variable. We can find the x-axis for these histograms at either the top or bottom of their respective column in the matrix and the y-axis at either the left or right of their respective row. The upper right corner of the matrix is filled with the correlation values between the various variables (for example, the value of 0.06 in the first row and the second column shows the level of correlation between age and gender. This helps us understand the relationships and any possible collinearity that could be occurring. An example of this would be the 0.06 value described above showing a low correlation between age and gender, or the value of 0.85 between occupation and income showing a high correlation. The bottom left corner of the matrix is filled with scatterplots that describe the relationship between each variable, as well as a simple linear regression line. The axes are the same as for the histograms.

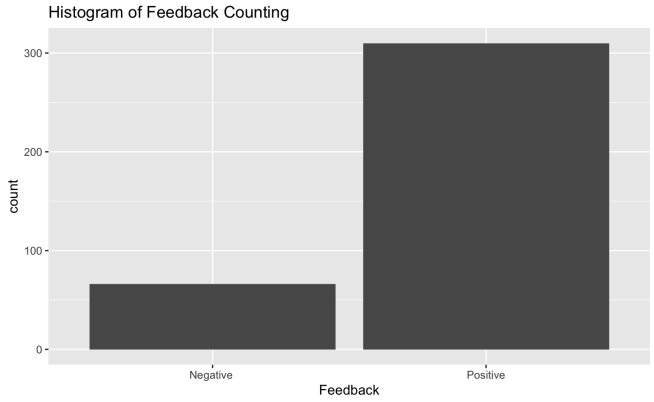


Above are box and whisker plots and bar charts that show the relationship between feedback and each predictor variable. The box plots are able to show us how the data of the predictor variable is distributed differently for negative and positive feedback. Starting from the left of the top row and moving right, we see that the 25th, 50th, and 75th percentiles of the age data for positive feedback are lower than those percentiles of the ages with negative feedback, indicating that in general positive feedback came from younger individuals. Next, we see that although the medians for the latitudes of the positive and negative groups are very similar, both the 25th and 75th percentiles are lower for the positive group, indicating that the typical positive feedback would be more likely to come from an individual from a location at a lower latitude. The opposite is true for longitude, where we see that positive feedback seems to be more likely to come from those living at a location with a higher longitude.

The bar plots are able to describe the ratio of positive vs. negative feedback for each level of those predictors. The marital status plot indicates that although the number of negative reviews was similar between married and single users, single users left far more positive reviews. Next, males left slightly higher negative and positive reviews, showing a similar ratio of positive to negative as female users. Our occupation plot shows our four different groups, which are all represented differently. Users with the occupation of housewife exclusively responded positively, self-employed users had about twice as many positive responses as negative responses, and employees had about three times as many positive reviews

as negative. Students had by far the largest discrepancy between positive and negative reviews, a useful finding from this plot that could prove useful in understanding a potential model. For our educational qualifications plot, the key finding was that those with Phds left far more positive reviews in comparison to negative reviews. Lastly, our income plot makes it clear that in this dataset, those with no income are far more likely to leave a positive review.

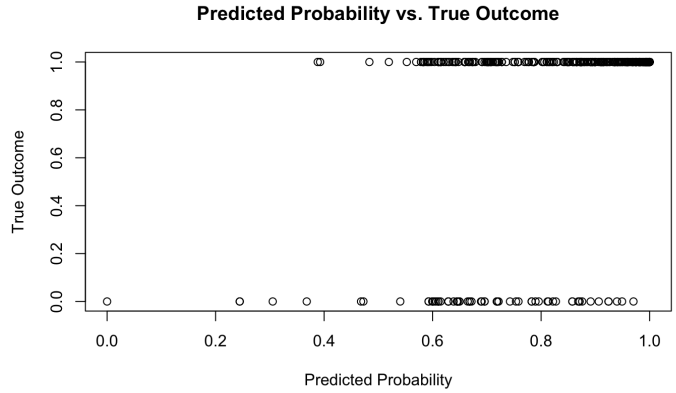
Using both of these types of plots is helpful to understand the relationship that each factor has on feedback, and how this may come into play in our models.



The bar chart just simply shows the distribution of our data response variable. As we can see, the number of positive feedbacks is much higher than the number of negative feedbacks. This suggests that the feedback in our dataset has been mostly positive.

V. Analysis

Logistic Regression Model:



In the plot above, we have our true outcomes for feedback (either 1 or 0, representing positive or negative) vs. the predicted probability of each of those points being positive under our logistic regression model. As we can see, our model gave the large majority of points across both classes high chances to be positive feedback, and many were, but there were still a decent amount of false positive predictions as some of the points had low predicted probability. The majority of the negative feedback points had high predicted probabilities, but there were also a couple of points with low predicted probabilities, showing some correct negative feedback predictions. Overall, both classes have high predicted probabilities of predicting a positive review.

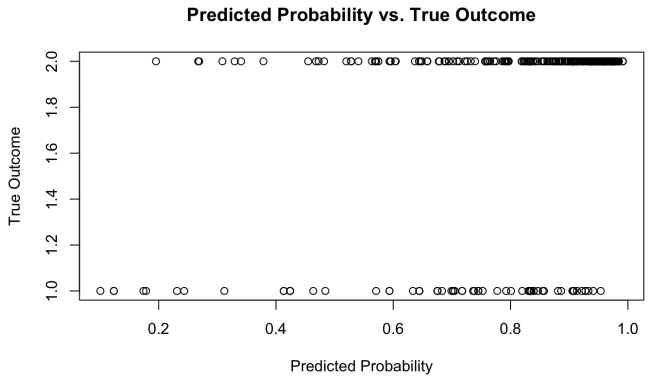
Confusion Matrix

Predicted/True	Negative	Positive
Negative	13	10
Postive	53	300

Upon examining the Confusion Matrix table provided, it is observed that the true positive rate, calculated as the ratio of correctly predicted positive instances to the total actual positives, $\frac{300}{10+300} = 0.967$, stands at 96.7%. Similarly, the positive predictive value, reflecting the proportion of true positive predictions among all positive predictions, is calculated as $\frac{300}{53+300} = 0.849$, which is 84.9%. These high percentages indicate a robust fit of the model to the data. The accuracy of the model, computed as the proportion of correctly classified instances among the total instances, is determined to be $\frac{13+300}{13+10+53+300} = 0.832$. The model achieves an accuracy of 83.2%, this level of accuracy suggests commendable overall performance in predicting feedback outcomes.

Utilizing logistic regression to analyze our dataset, we regress the response variable (Feedback) against all other predictor variables. The resulting Area Under the Receiver Operating Characteristic Curve (AUC) stands at 0.788, indicating a substantial predictive accuracy. This robust performance underscores the model's efficacy in discerning patterns and relationships within the data, thus offering valuable insights into the factors influencing feedback outcomes.

LDA Model:



Above we have a similar plot to the one previously described in the logistic regression section, showing the LDA model's predicted probability of each data point being a positive review (true outcome = 2), or a negative review (true outcome = 1). We can see that the points are more spread out in terms of their probabilities, giving more points low chances to be true outcomes, across both classes than in the logistic regression plot of the same nature.

Confusion Matrix

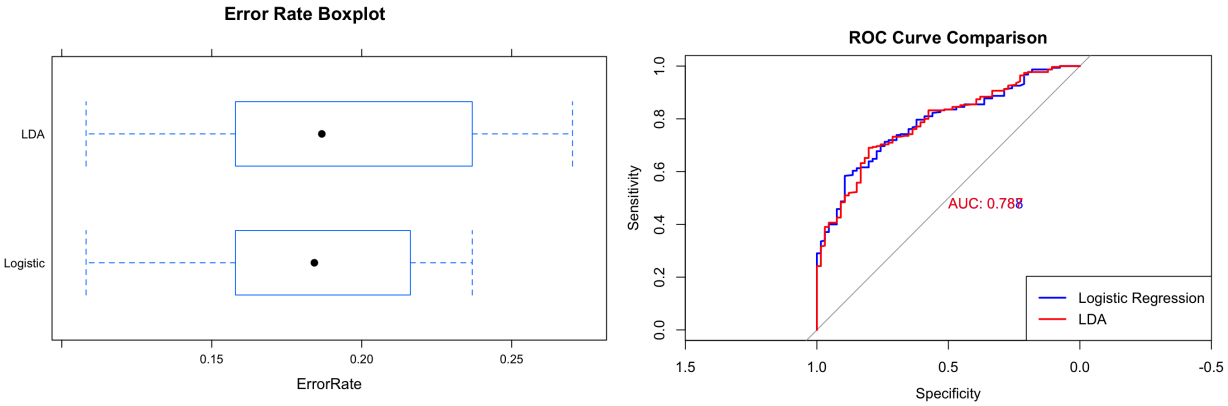
Predicted/True	Negative	Positive
Negative	15	11
Postive	51	299

Upon analyzing the provided Confusion Matrix table, it becomes apparent that the true positive rate, $\frac{299}{11+299} = 0.964$, indicative of the model's ability to correctly identify positive instances, stands notably high at 0.964. Similarly, the positive predictive value, $\frac{299}{51+299} = 0.854$, reflecting the accuracy of positive predictions made by the model, is calculated at 0.854. These metrics collectively suggest a strong alignment between the model's predictions and the actual outcomes. Moreover, the model demonstrates an accuracy rate calculated as $\frac{15+299}{15+11+51+299} = 0.835$. 83.52% accuracy rate reflects the model's capability to accurately classify instances across all categories. This level of accuracy underscores the model's robust performance and its effectiveness in capturing the nuances of feedback outcomes.

Utilizing Linear Discriminant Analysis (LDA) to analyze our dataset, we regress the response variable (Feedback) against all other predictor variables. The resulting Area Under the Receiver Operating Characteristic Curve (AUC) is calculated to be 0.787, denoting a notable degree of predictive accuracy.

This finding underscores the effectiveness of the LDA model in discerning patterns and relationships within the data, thus providing valuable insights into the factors influencing feedback outcomes.

Cross-Validation



When employing the Cross-Validation technique, the computed mean error rates serve as illuminating metrics for evaluating the performance of both the logistic regression and Linear Discriminant Analysis (LDA) models. Specifically, the logistic regression model demonstrates a commendable mean error rate of 0.1834282. This figure provides valuable insights into the accuracy and effectiveness of the logistic regression model in predicting user feedback on the online food ordering experience. Conversely, the LDA model yields a slightly higher mean error rate of 0.1967994. This comparison suggests that, while both models offer predictive capabilities, the logistic regression model may outperform the LDA model in terms of overall accuracy.

After a comprehensive analysis of the logistic regression and Linear Discriminant Analysis (LDA) models, we have reached the decision to favor the logistic regression model for predicting user feedback on the online food ordering experience. The logistic regression model demonstrates robust performance across various evaluation metrics. It achieves a true positive rate of 96.7% and a positive predictive value of 84.9%, indicating high accuracy in identifying positive instances and making positive predictions. Moreover, the model exhibits an impressive accuracy rate of 83.2%, highlighting its capability to correctly classify instances across all categories. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC) for the logistic regression model stands at 0.788, further underscoring its substantial predictive accuracy. While both models offer predictive capabilities, the logistic regression model consistently outperforms the LDA model across multiple metrics, suggesting superior overall accuracy and effectiveness in capturing the nuances of feedback outcomes. Therefore, based on these

findings, we conclude that the logistic regression model is the preferred choice for our predictive modeling task.

VI. Result

Based on our analysis and consideration of several variables we can infer that predictors which include gender, marital status, age, occupation, monthly income, education, family size, latitude, and longitude are essential variables and have an impact on the tendency of feedback being either positive or negative. Gender being male, an increase in age, and an increasing family are all associated with a higher likelihood of the feedback being positive. On the other hand, marital status of being single, occupations such as housewife, self-employed, and student, and an increase in monthly income and qualification are associated with a higher likelihood of the feedback being negative.

Through our analysis, we determined that the Pin.code variable does not significantly impact a customer's rating tendencies. Additionally, rows where the marital status is listed as "prefer not to say" were excluded to ensure reliable and representative data. All other variables were found to be significant for creating our predictive models. We performed model selection within each method, and the remaining variables were all significant contributors to our final models.

Based on our analysis both the models obtained from logistic regression and LDA would produce reasonable classifications of the feedback when introduced with new datasets with customer information. However, we would further consider the logistic regression model because it provides more accurate and actionable insights into user preferences and the factors influencing their continued use of online food ordering platforms. It achieves an accuracy of 83.2%, with a true positive rate of 96.7% and a positive predictive value of 84.9%. The AUC for logistic regression is 0.788, indicating substantial predictive accuracy. Compared to the LDA model, which has a slightly higher accuracy (83.5%) but a higher mean error rate (0.1968 vs. 0.1834 for logistic regression), logistic regression consistently performs better across various metrics, suggesting superior overall accuracy and effectiveness in capturing the nuances of feedback outcomes.

VII. Conclusion

In sum, we used the techniques of both logistic regression and linear discriminant analysis (LDA) to create models that could accurately predict, based on demographic and geographic data, whether a customer's review for an online food order would be negative or positive. Both models were successful in

this task, but our logistic regression model proved to be better suited as it outperformed the LDA on several evaluation metrics, including accuracy, mean error rate, AUC, true positive rate, and positive predictive value.

Further analysis of our logistic regression model provided insights into which categories within our predictors influenced the likelihood of a positive or negative review. For instance, male users, older individuals, and those with larger family sizes are more likely to leave positive feedback. Conversely, single users, those with higher incomes, higher education levels, and certain occupations such as housewives, self-employed individuals, and students are more likely to leave negative feedback.

Any company that receives reviews would prefer to have a high positive-to-negative review ratio, as this can often be seen by consumers through online review boards. Consumers are more likely to use a service that is highly rated, which in turn can increase revenue for the company. Therefore, this analysis could be invaluable to online food delivery services keen to improve their services and increase revenue. By understanding which groups are less likely to leave positive reviews, these services can tailor their marketing or customer service strategies to better address and meet the needs of these specific customer segments.