**MACHINE LEARNING PS-3**

2. ANALYSIS

(e) Coefficients Table

Model Intercept Review\_Scores\_Rating

1 Linear Probability Model - 1.217396 0.3215106

2 Logit Model -26.008035 5.2137464

3 Probit Model -16.286638        3.2655207

Below is the interpretation of the coefficients from the three models:

***Linear Probability Model:***

* Intercept: -1.217396: This indicates that when the review score is 0, the predicted probability of being a superhost is approximately -121.74%. Since probabilities cannot be negative, this result highlights a limitation of the Linear Probability Model (LPM).
* Review Scores Rating: 0.3215106: For each one-point increase in the review score, the probability of being a superhost increases by about 32.15%. This means that higher review scores positively influence the likelihood of being classified as a superhost.

***Logit Model:***

* Intercept: -26.008035: This represents the log-odds of being a superhost when the review score is 0. Like the LPM, this value does not have a straightforward interpretation in terms of probability but serves as a baseline for calculating odds.
* Review Scores Rating: 5.2137464: For each one-point increase in the review score, the log-odds of being a superhost increase by 5.21. This means that for each one-point increase in the review score, the odds of being a superhost increase by about 182.48 times, highlighting a strong positive relationship between review scores and the likelihood of becoming a superhost.

***Probit Model:***

* Intercept: -16.286638: Similar to the logit model, this value represents the starting point when the review score is 0, but it is not directly interpretable as a probability.
* Review Scores Rating: 3.2655207: A one-point increase in the review score raises the probability of being a superhost, following a standard normal distribution. While the exact probability increase is not as straightforward as in the LPM, it suggests that higher review scores significantly improve the likelihood of being classified as a superhost.

***Causal Interpretation:***

The analysis suggests a causal relationship between review scores and the probability of being classified as a superhost, as indicated by the consistent positive coefficients across the Linear Probability Model, Logit Model, and Probit Model. Specifically, each one-point increase in review scores is associated with a significant increase in the likelihood of achieving superhost status, with the Linear Probability Model estimating a 32.15% increase in probability. However, to confidently assert causality, it is crucial to establish that review scores precede superhost classification and to control for potential confounding variables that could bias the results. Overall, the findings support the inference that higher review scores lead to an increased likelihood of becoming a superhost, assuming the underlying assumptions for causal inference are satisfied.

(f) Based on the cross-validation results, the error rate is the same (0.2695) across all tested cost parameters {10^-2, 10^-1, 10^0, 10^1, 10^2}. Since there is no significant difference in the error, the optimal cost parameter can be chosen based on other considerations, such as simplicity or efficiency. In this case, the smallest cost value, 0.01 (or 1e-02), would be a reasonable choice, as it provides a more regularized model without increasing the error.

(g)

(svm\_model\_ performances)

gamma cost error dispersion

1 0.1 1e-02 0.2700963 0.006524819

2 0.1 1e-01 0.2700963 0.006524819

3 0.1 1e+00 0.2700963 0.006524819

4 0.1 1e+01 0.2700963 0.006524819

5 0.1 1e+02 0.2700963 0.006524819

Intuitively, A low value (e.g., 0.1) leads to a smoother, more generalized decision boundary, which might not capture complex patterns. A high value (e.g., 10) makes the model more sensitive to individual data points, creating a tighter boundary but increasing the risk of overfitting. Thus, a higher value improves accuracy on the training set but risks poor generalization on new data.

In my tuning process, I found that the optimal cost parameter for this gamma value is 100, indicating that this setting may balance complexity and generalization effectively for my dataset.

(i) Model Mean Classification Error

1 Linear Probability Model 0.267

2 Logit Model 0.259

3 Probit Model 0.261

4 SVM ( = 0.1) 0.689

5 SVM ( = 10) 0.583

6 Lasso Logistic 0.194

The mean classification errors indicate the performance of different models in predicting superhost status. The Lasso Logistic model has the lowest error at 0.194, suggesting it is the most accurate in this context. The Logit model follows with an error of 0.259, while the Probit model is slightly worse at 0.261. The Linear Probability Model has a higher error of 0.267. In contrast, the SVM models show significantly higher errors, particularly with =0.1 at 0.689, indicating poorer predictive performance compared to the other models.

2. CONCEPTUAL PROBLEMS

(a) In the Multinomial Logit (MNL) model, each consumer's choice of an iPhone model is based on the utility they get from the different options. Given that price is the only factor influencing the choice, and the effect of price is the same for all models (represented by the coefficient ), the probability of a consumer choosing the iPhone 16 is:

/ }

where β is the price sensitivity coefficient, and represents the price of each iPhone model (e.g., iPhone SE, iPhone 15, iPhone 16, iPhone 16 Pro). The numerator gives the utility from iPhone 16, and the denominator sums the utilities for all models, reflecting consumer choice as a function of price comparisons.

This formula shows that the probability of choosing the iPhone 16 depends on its price relative to the other models. If the iPhone 16 has a lower price compared to the other models, its probability of being chosen increases. The MNL model captures the trade-offs consumers make between different product options, and the probability is influenced by the prices of all available iPhone models.

(b) To predict the market share of the new "iPhone 16 Super" using the estimated Multinomial Logit (MNL) model, we calculate its probability of being chosen relative to the other iPhone models. The formula is:

To estimate this, we would assign a price to the iPhone 16 Super and include it in the calculation along with the prices of the existing models (iPhone SE, iPhone 15, iPhone 16, and iPhone 16 Pro). The predicted market share depends on how its price compares to the other models—lower prices or better features will increase its predicted share.

(c) The MNL model predicts how the demand for each iPhone model changes based on prices. The demand for the iPhone 16 (or any model) is influenced by its price relative to the other models. This allows us to estimate how pricing changes affect market shares and ultimately revenue. For the new iPhone 16 Super, we can predict its market share by comparing its price to other models, guiding our pricing strategy to maximize sales and profit.

(d) When a new iPhone model (iPhone 16 Super) is introduced, the market shares of the existing models must decrease. This is because the sum of the probabilities must still equal 1, and adding a new model divides the total probability between more options. As a result, the iPhone 16 Super will take market share from other models. Companies can use this insight to adjust their marketing strategies and production plans to maximize overall profitability while accounting for the potential loss in market share for older models.