Customer Purchase Prediction

Yuxuan Zhang

Westcliff University

AIT 506

Instructor: Professor Noha Bhairy

Date: Nov 15, 2025

Customer Purchase Prediction

The goal of this project is two fold. First, we aim to understand how different customer attributes – such as age and income – impact whether a purchase will be made. More precisely, we want to understand the type of influence of each factor on the purchase decision as well as the magnitude of that influence. From this understanding, we look to come up with marketing strategies for relevant organizations to improve sales and revenue. Secondly, we want to build machine learning models that can predict the customer purchase decision with good performance.

The dataset consists of customer attributes such as the age and education level of the customer as well as product attributes such as the review type (good, average, bad). In an initial data analysis, we plotted the relationships between each of the factors and the target variable. From the graphs, the following observations can be made.

1. A higher age is associated with a higher rate of purchase.
2. The “school” education level is associated with a lower rate of purchase.
3. The “average” review type is associated with a lower rate of purchase.

Since these observations are drawn from visualizations of a single parameter versus the target variable and do not take into account the correlations with other variables, they only provide a preliminary peek into the data and caution must be taken when interpreting them.

Before training models for prediction and interpretation, some preprocessing steps are taken on the data beforehand. The review column is turned into an ordinal variable because the review type naturally has an ordering from low to high. The gender and education level are one hot encoded. Finally, Z score normalization was performed on all numerical variables (de Amorim, 2023). A train-test split of 80-20 was taken.

The models that are experimented with are logistic regression, decision tree and XGBoost. For all models, cross validation is used to find the best hyperparameters (Geisser, 1975). For example, for logistic regression, the hyperparameters that are tuned are the regularization parameter and the regularization type. For decision trees, the hyperparameters that are tuned include tree depth, the cost complexity parameter and the type of loss used in the growing step of the algorithm.

Overall, logistic regression performs the best with an F1 score of 0.6. By visualizing the logit coefficients, we make the following observations.

1. Education_school has the highest impact on purchase decisions and it decreases the chance that a purchase will be made. This agrees with what we observed from the graphs in the initial visualization analysis.
2. Among all variables that have a positive impact on purchase, Education_UG has the largest coefficient. This means that being on the undergraduate education level makes it more likely that the customer will make the purchase, when all other variables are fixed.

The decision tree has an F1 score of 0.57. By visualizing the tree, we see the same types of impact as drawn from the logistic regression model. For example, in the split on the right on the second level, education_school leads to a leaf node for not making the purchase. Feature importance – measured by the total reduction in node impurity caused by each variable (Breiman, 1984), however, suggests that age has the greatest impact on the purchase decision. The full feature importance graph can be viewed in the notebook.

Finally, from the observations we make from model interpretation as well as the exploratory data analysis, we propose a few strategies to the company for improving sales and revenue. First, the company may consider allocating more marketing resources to customers

with education levels strictly higher than school and belonging to higher age groups. Also, it might be reasonable to generate positive reviews on products with a current average review score.

References

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Wadsworth.

de Amorim, L. B. V., Cavalcanti, G. D. C., & Cruz, R. M. O. (2023). The choice of scaling technique matters for classification performance. Applied Soft Computing, 133, 109924. https://doi.org/10.1016/j.asoc.2022.109924

Geisser, S. (1975). The predictive sample reuse method with applications. Journal of the American Statistical Association, 70(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865