STAT 627 Statistical Machine Learning Project

Project Name:

Prediction and analysis of in-vehicle coupon acceptance behavior

Class Section: STAT-627-002

Team Members: Shu-Wen Teng, Yen-Jo Lee

Last updated: 2024/12/04

Description of the Dataset

The dataset for this project is sourced from the UC Irvine Machine Learning Repository and was donated in 2020. It was collected through a survey conducted on Amazon Mechanical Turk and is referenced in the paper *"A Bayesian Framework for Learning Rule Sets for Interpretable Classification"* by Wang et al. (2017). The dataset explores factors influencing whether a person will accept a coupon recommendation in various driving scenarios. It comprises 12,684 observations across 25 variables. These variables capture a variety of details, including:

- Driver Demographics: Age, marital status, and whether the driver has children.
- Financial Information: Income levels.
- Contextual and Marketing Interaction Details: Driving scenarios, such as destination, time, weather, and passenger type, as well as coupon-related details, including type, expiration time, and driving distance to the location where the coupon can be redeemed.

Research Question

Which machine learning model best predicts whether a driver will accept an in-vehicle coupon based on various contextual factors?

Method

We cleaned the dataset by first removing specific columns that were deemed unnecessary for further analysis, such as occupation, car details, and various dining preferences. This step helps streamline the dataset by focusing on attributes relevant to our analysis. Next, we transformed categorical variables such as 'destination', 'weather', 'time', 'coupon', 'expiration', 'gender', and 'maritalStatus' into numerical values using a label encoder. This transformation is critical because machine learning algorithms typically require numerical inputs.

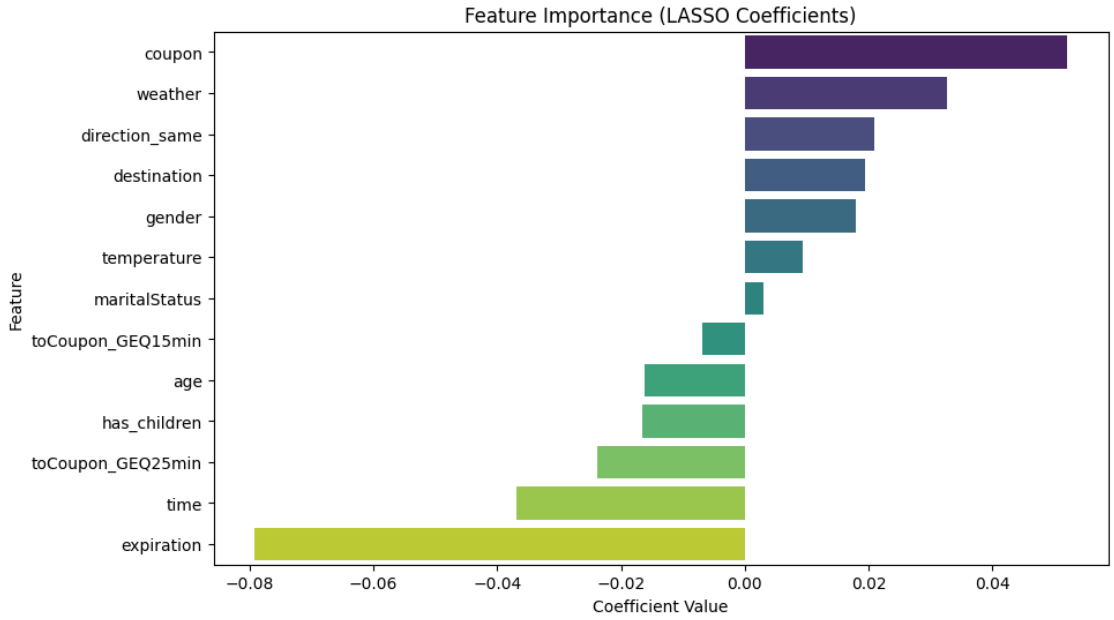
To evaluate the cleaned dataset, we applied five machine learning methods to predict the target variable:

- LASSO: Used for feature selection, identifying the most relevant variables by penalizing less important ones.
- Logistic Regression: A statistical method for binary classification tasks, used as a baseline model.
- K-Nearest Neighbors: A simple, instance-based learning method that predicts based on the proximity of training data points.
- Linear Discriminant Analysis: A method that maximizes class separability for classification tasks.

- Decision Trees: A versatile algorithm that splits data based on feature values to make predictions.
- Support Vector Machines: A robust algorithm that finds the optimal hyperplane for classification.

We split the dataset into training 70% and testing 30% sets to evaluate the performance of each model. All analyses and modeling steps were conducted using Python in Google Colab.

LASSO



After cleaning the data, which still contained numerous predictors, we utilized LASSO regression for both variable selection and regularization, using 5-fold cross-validation. Interestingly, the model retained all predictors, indicating that each variable contributes in some way to the overall prediction. This approach improves the model's predictive accuracy and interpretability by shrinking the coefficients of less important variables toward zero. The table of non-zero coefficients highlight the features that significantly contribute to predicting the target variable.

The magnitude of the coefficients provides insights into each feature's impact. For instance, expiration (-0.079247) and time (-0.036868) have the strongest effects (both negative). Coupons with shorter expiration times are more likely to be accepted, which is probably because it makes people feel more urgent. While demographic factors like age (-0.016215) and has_children (-0.016643) exert smaller influences.

Logistic Regression

```
Optimization terminated successfully.
         Current function value: 0.656566
         Iterations 5
```

Logit Regression Results

Dep. Variable:	Y	No. Observations:	12684
Model:	Logit	Df Residuals:	12671
Method:	MLE	Df Model:	12
Date:	Sun, 01 Dec 2024	Pseudo R-squ.:	0.03976
Time:	22:20:34	Log-Likelihood:	-8327.9
converged:	True	LL-Null:	-8672.7
Covariance Type:	nonrobust	LLR p-value:	7.234e-140

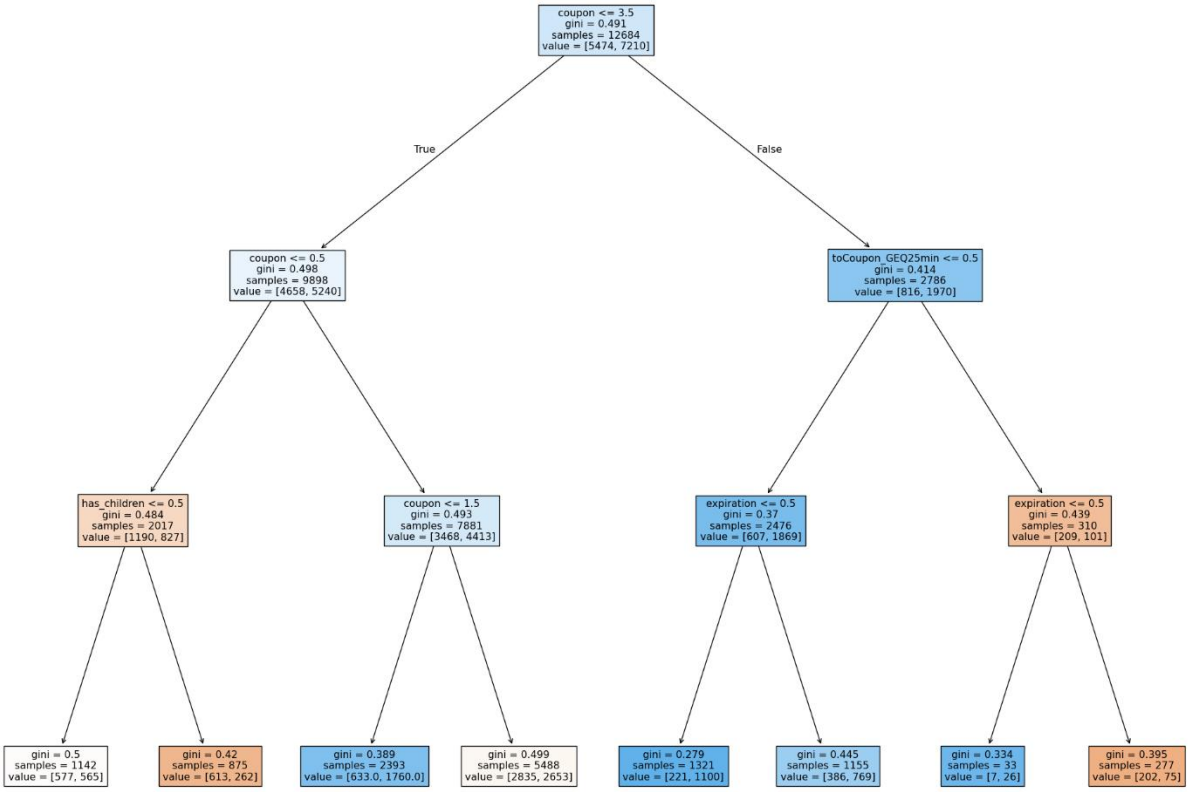
	coef	std err	z	P>|z|	[0.025	0.975]
destination	0.1421	0.029	4.862	0.000	0.085	0.199
weather	0.2166	0.031	6.878	0.000	0.155	0.278
temperature	0.0013	0.001	1.353	0.176	-0.001	0.003
time	-0.1171	0.016	-7.160	0.000	-0.149	-0.085
coupon	0.1645	0.014	11.499	0.000	0.136	0.193
expiration	-0.6573	0.038	-17.131	0.000	-0.732	-0.582
gender	0.1607	0.036	4.413	0.000	0.089	0.232
age	-0.0216	0.009	-2.540	0.011	-0.038	-0.005
maritalStatus	0.0062	0.021	0.291	0.771	-0.036	0.048
has_children	-0.1163	0.042	-2.800	0.005	-0.198	-0.035
toCoupon_GEQ15min	-0.1246	0.040	-3.122	0.002	-0.203	-0.046
toCoupon_GEQ25min	-0.3263	0.066	-4.964	0.000	-0.455	-0.197
direction_same	0.1721	0.053	3.225	0.001	0.068	0.277

Our analysis identified several significant variables, including destination, weather, time, coupon, expiration, and gender, which contribute meaningfully to our model. In contrast, temperature ($p = 0.176$) and marital status ($p = 0.771$) were found to be statistically insignificant, indicating that these variables have little impact on predicting coupon acceptance.

Decision Tree

Through 10-fold cross-validation, we have identified the optimal hyperparameters for the decision tree model: a maximum depth of 10, a minimum of 1 sample per leaf, and a minimum of 2 samples required to split an internal node. These settings suggest that the model can capture a moderate level of complexity while avoiding overfitting.

This decision tree diagram provides an insight into the factors influencing drivers' acceptance of in-vehicle coupons. The root node uses the 'coupon' attribute for initial splitting, indicating its primary importance in the decision-making process. The tree splits further into various nodes based on other attributes such as 'toCoupon_GEQ25min' (time to coupon), 'expiration', and 'has_children', highlighting their roles in coupon acceptance. Lower Gini impurity values at a node suggest higher homogeneity, implying better classification at that decision point.



K-Nearest Neighbors

We use the K-Nearest Neighbors classifier for prediction, achieving about 63.48% accuracy on the test dataset. The classification report shows that the model has precisions of 0.59 for class 0 and 0.66 for class 1, with recall rates of 0.52 and 0.72 respectively, and F1-scores of 0.55 and 0.69. These metrics suggest that the model performs better at identifying class 1. Additionally, 10-fold cross-validation was conducted, with a mean MSE of approximately 0.349.

Linear Discriminant Analysis

In the LDA model, we achieved a test dataset accuracy of about 61.42%. The model was more effective in identifying class 1 (precision 0.63, recall 0.80) than class 0. The confusion matrix and a mean squared error of 0.3816 from 10-fold cross-validation suggest moderate prediction errors. Utilizing GridSearchCV, the best LDA parameters found were 'shrinkage: None' and 'solver: lsqr', resulting in a slight improvement in cross-validation accuracy to 0.6188. This indicates that parameter optimization can enhance model performance.

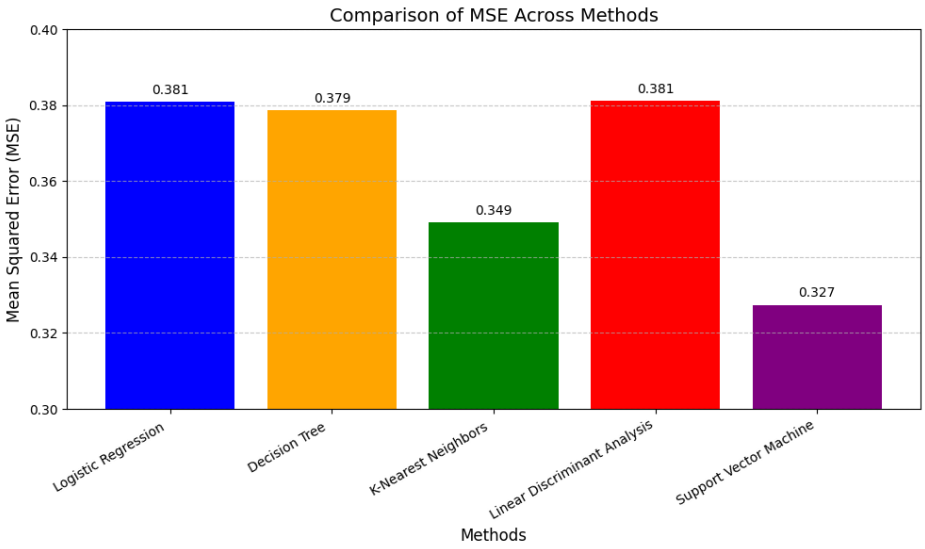
Support Vector Machine

For Hyperparameter Tuning, we performed a grid search using GridSearchCV to optimize the parameters of an SVM model with an RBF kernel. The results show that the best cross-validation score of 0.67458 was achieved with a regularization parameter C=10 and γ=0.1. Additionally, 10-fold cross-validation was conducted, with a mean MSE of approximately 0.327.

Result and Conclusion

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	61.56	61.0	59.0	60.0
Decision Tree	62.01	62.0	62.0	62.0
K-Nearest Neighbors	63.48	63.0	62.0	62.0
Linear Discriminant Analysis	61.43	60.0	59.0	58.0
Support Vector Machine	65.92	66.0	64.0	65.0

- Values are rounded to one decimal place for percentages.
- **Precision**, **Recall**, and **F1-Score** are averaged across classes (weighted average).



After reviewing the performance of all the classification models, we conducted support vector machine performs well overall. Compared to models like Logistic Regression, Decision Tree, K-Nearest Neighbors, and Linear Discriminant Analysis, SVM model is the best in accuracy, precision, recall, F1-score, and lowest MSE. The Decision Tree and K-Nearest Neighbors models also showed relatively good results. In contrast, Logistic Regression and Linear Discriminant Analysis were performed modestly.

In conclusion, our analysis concludes that a Support Vector Machine is the best way to predict whether drivers will accept in-vehicle coupons. However, this study is limited by its reliance on a single dataset collected via surveys, which may not fully capture real-world behaviors. We recommend that Amazon Mechanical Turk could explore additional datasets in the future, integrate real-time data, or test advanced algorithms like ensemble methods to validate and improve upon these results.

The contribution of each team member:

We did most of the code together and did the explanation separately.

Shu-Wen Teng: Decision tree, Linear Discriminant Analysis, KNN

Yen-Jo Lee: LASSO, Logistic Regression, SVM

References

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2015). Or's of And's for Interpretable Classification, with Application to Context-Aware Recommender Systems. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1504.07614

N, R. R., Jain, S., & Sarkar, M. (2023). SMOTE and Hyperparameter Optimization: a dual machine learning strategy for enhancing coupon recommendation in vehicular contexts. *2023 3rd International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*, 1–6. https://doi.org/10.1109/smartgencon60755.2023.10442306