

**CIND 820 Capstone Project Final Report: In-Vehicle Coupon Recommendation**

Sucheta Sikdar (ID: 501074722)

Supervisor: Dr. Ceni Babaoglu

Toronto Metropolitan University

Date of Submission: 2024-11-30

Table of Contents

[Abstract 3](#_Toc183859420)

[Introduction 7](#_Toc183859421)

[Literature Review 9](#_Toc183859422)

[Data Description 16](#_Toc183859423)

[Dimensionality Reduction 41](#_Toc183859424)

[Classification Algorithms and Cross Validation 49](#_Toc183859425)

[Comparison and Analysis of results 56](#_Toc183859426)

[Methodology 58](#_Toc183859427)

[Limitations / Challenges / Continuity 60](#_Toc183859428)

[GitHub Repository 61](#_Toc183859429)

[References 61](#_Toc183859430)

# Abstract

Coupons are a great way for customers to save money on their purchases. It is also a great way for business to attract customers. Coupons create a win-win situation for both companies and customers and hence by offering the correct coupon to users can lead to users to become frequent customers (Niralidedaniya, 2023). If businesses can find the right customers who will use their coupons, then it will help businesses. It will also be interesting to predict what type of coupon will be accepted by a customer based on various attributes about the customer. Luckily, there exists a publicly available dataset called In-Vehicle Coupon Recommendation at UCI Machine Learning Repository from 2020 which describes different driving scenarios of multiple clients such as destination, time, coupon, expiration, gender, age, marital status, whether they have children, education, occupation, income, car, the number of times they go to the bar per month, the number of times they go to a coffee shop per month, the number of times that they buy take away food per month, if customer’s average expense per person at restaurants is less than 20 dollars a month, if customer’s average expense per person at a restaurant is between 20 dollars to 50 dollars per month, driving distance to the restaurant/bar for using the coupon is greater than 15 minutes, driving distance to the restaurant/bar for using the coupon is greater than 25 minutes, whether the restaurant/bar is in the same direction as destination, whether the restaurant/bar is in the opposite direction as destination, whether the coupon is accepted (*UCI Machine Learning Repository*, 2020).

When printing the info of the dataset in VSCode using Python, the following was observed:

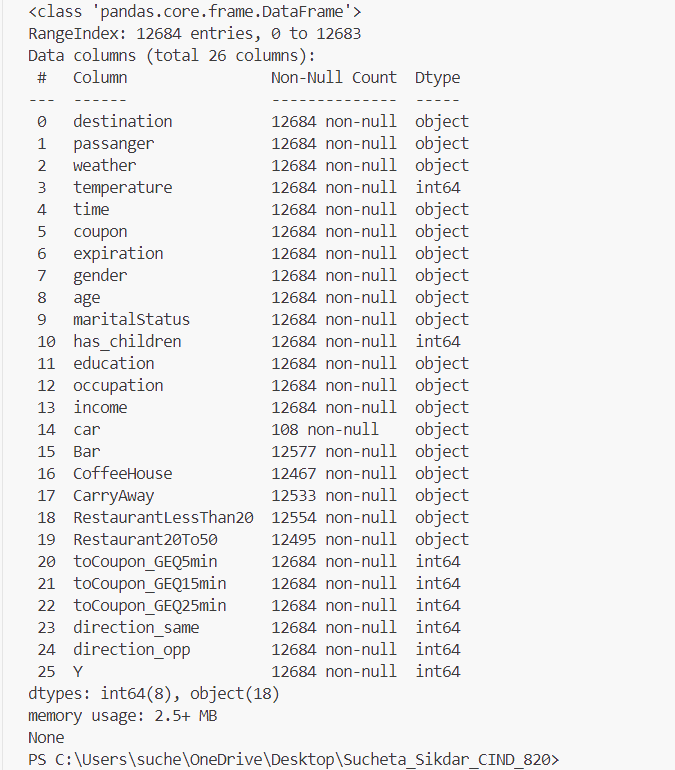


Figure 1: Info on the In-Vehicle Coupon Recommendation Dataset

As seen in Figure 1, the dataset has 12684 records. The dataset has 25 features and 1 target column. Some of the data like marital status and gender are categorical and some of the data like age and temperature are numerical. Using code provided by Niralidedaniya (2023), it was found that the target classes are partially balanced. If the target classes were highly unbalanced, then this dataset could not be used because the results of supervised learning algorithms used to make predictions would skew towards the class with the class with higher percentage of records.

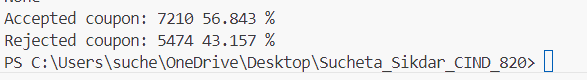


Figure 2: Distribution of Target Classes

It should also be noted that this dataset comes with many missing values. Hence, this dataset requires preprocessing before it can be analyzed with machine learning algorithms. Using the code provided by Niralidedaniya (2023), we can see the features which have missing values.

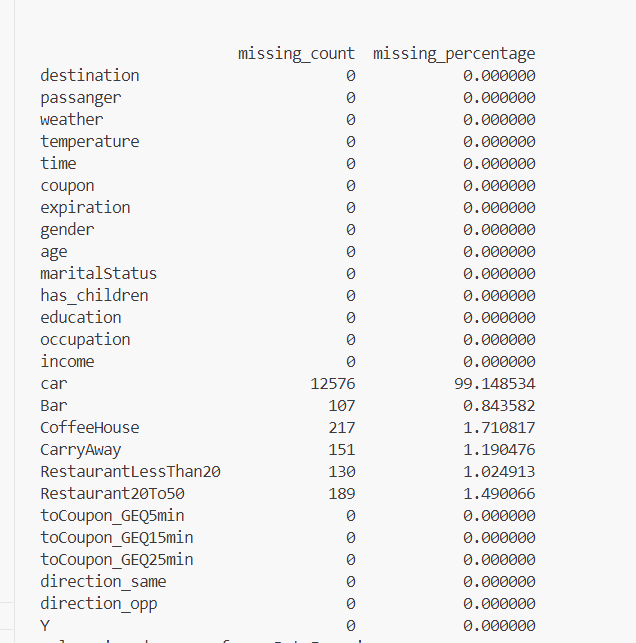


Figure 3: Distribution of missing values in the dataset

The objective of this project will be to:

* Find the best predictive classification algorithm for the In-Vehicle Coupon Recommendation dataset (2020) after evaluation of various supervised learning classification algorithms introduced to us in *CMTH 642 – Data Analytics: Advanced Methods* like Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-Nearest Neighbours (k-NN) on the dataset.
* Using a correlation matrix find out which attributes are highly correlated to the target of the customer accepting or rejecting a coupon.
* Find whether we can attain a dataset with fewer dimensions using these 3 methods: Stepwise Regression, Forward Feature Selection and Backward Feature Elimination methods learnt by us in *CMTH 642 – Data Analytics: Advanced Methods*.
* Find the limitations of this dataset.

Data analysis for this project will be done using Python. Pandas and numpy libraries will be used. Seaborn and Matplotlib libraries will be used for visualizations. Supervised learning techniques will require the scikit-learn library (*Supervised Learning*, n.d.). Evaluation of various classification algorithms will be done by comparing evaluation metrics like the Accuracy and Area under the Curve (AUC) of each algorithm.

# Introduction

Coupons are a great way for customers to save money on their purchases and feel special that they are getting a discount. Businesses can attract customers with coupons. If businesses can find the right customers who will use their coupons, then it will help businesses survive and grow. Businesses can then retain existing customers and attract new customers. If businesses know beforehand as to which customers to target for their coupons, then it will save them both money and effort in marketing and sending coupons (Ahmed et al., 2024).

It will also be interesting to predict what type of coupon will be accepted by a customer based on various attributes about the customer. There exists a publicly available dataset called In-Vehicle Coupon Recommendation at UCI Machine Learning Repository from 2020 which describes different driving scenarios of multiple clients and whether the coupon is accepted (*UCI Machine Learning Repository*, 2020).

This project will use the knowledge that has been gained in previous courses of the Data Analytics, Big Data, and Predictive Analytics Certificate Program taught at Toronto Metropolitan University. What has been learned can be implemented in this project.

The objective of this project will be to:

* Find the best predictive classification algorithm for the In-Vehicle Coupon Recommendation dataset (2020) after evaluation of various supervised learning classification algorithms introduced to us in *CMTH 642 – Data Analytics: Advanced Methods* like Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-Nearest Neighbours (k-NN) on the dataset. What has been learned will be implemented.
* Using a correlation matrix find out which attributes are highly correlated to the target of the customer accepting or rejecting a coupon. A visual of a correlation matrix will be very effective. **Since most of the features are categorical, the discreet values will be converted to corresponding numerical values using encoding.**
* Find whether we can attain a dataset with fewer dimensions using these 3 methods: Stepwise Regression, Forward Feature Selection and Backward Feature Elimination methods learnt by us in *CMTH 642 – Data Analytics: Advanced Methods*. This will help eliminate some noise. **Dimensionality reduction is the major contribution of this project to this dataset.**
* Find the limitations of this dataset. The flaws of this dataset need to be explored.

There is some existing and related work done on this topic already. Exploratory data analysis to get a better picture of our data is also necessary.

# Literature Review

The original data was collected by Wang et al. (2017) using Amazon Mechanical Turk about users interacting with a mobile recommendation system. They used rule set models that are used for classification and decision making to understand and predict consumers’ response to coupons in different contexts (Wang et al., 2017). Wang et al. (2017) used Bayesian approach and Disjunctive Normal Form classifiers. An example is that if a combination of a subset of input is met, then the output is satisfied. Wang et al. (2017) compared performance in accuracy by calculating and comparing Area under the ROC curve for their deduced final two Bayesian Rule Sets with other machine learning methods including different types of decision tree algorithms. Wang et al. (2017) compared with C4.5, CART, Lasso, RIPPER and TopK. Unlike other machine learning algorithms, a Bayesian approach looks at previous choices. Wang et al. (2017) deduced that that their Bayesian approach had competitive performance. Their methods are more complex compared to the simpler supervised learning classification algorithms introduced to us in *CMTH 642 – Data Analytics: Advanced Methods.* It is worth checking how this dataset performs against simpler supervised learning classification algorithms introduced to us in *CMTH 642 – Data Analytics: Advanced Methods.* It is possible that using simpler supervised learning classification algorithms like Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-Nearest Neighbours (k-NN) might yield similar competitive results too.

Same dataset was used by Niralidedaniya (2023). A lot of the data understanding, Data preparation and Exploratory Data Analysis on the dataset was already done by Niralidedaniya (2023). Niralidedaniya (2023) experimented with below thirteen machine learning classification models with hyper parameter tuning:

1. Logistic Regression
2. K-Nearest Neighbor
3. Decision Tree
4. Support Vector Classification with rbf kernel
5. LinearSVC
6. Gaussian Naive Bayes
7. Random Forest
8. GBDT
9. Bagging Classifier
10. AdaBoost Classifier
11. Gradient Boosting Classifier
12. ExtraTrees Classifier
13. HistGradientBoosting Classifier

Niralidedaniya (2023) compared the test log loss and test AUC score of each of the thirteen ML models with different encoded data matrices to find the best model and best encoding method with their best hyper parameter. Niralidedaniya (2023) found that Ordinal Encoding and One Hot Encoding perform best than other encoding techniques. It was observed that XGB Classifier, Support Vector Classification, Hist Gradient Boosting, and Random Forest Classifier models perform best than other models. Bagging, AdaBoost, Gradient Boosting, and ExtraTrees Classifier models perform best but they are overfitted (Niralidedaniya, 2023). Niralidedaniya (2023) stated that their Training AUC Score is very high with a value of 1, and a training log loss is almost 0. Basic classification models like Logistic Regression, K-Nearest Neighbor, Decision Tree, LinearSVC, and Gaussian Naive Bayes models didn’t do well for this problem (Niralidedaniya, 2023). Niralidedaniya (2023) took classification up a notch by the use of stacking classifier on the four models: XGB Classifier, Support Vector Classification, Hist Gradient Boosting, and Random Forest Classifier with their best parameters. It would still be interesting to see and compare how the basic classification algorithms introduced to us in *CMTH 642 – Data Analytics: Advanced Methods* like Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-Nearest Neighbours (k-NN) perform on the dataset. A generated visual correlation matrix will be insightful. A dataset with fewer dimensions can be attainted using these 3 methods: Stepwise Regression, Forward Feature Selection and Backward Feature Elimination methods learnt by us in *CMTH 642 – Data Analytics: Advanced Methods*. It will also be useful to explore the limitations of this dataset.

Depari et al. (2022) used the same dataset. Depari et al. (2022) used RapidMiner 9.10.001 tool in predicting customer’s responses to in-vehicle coupon recommendations. Depari et al. (2022) compared the accuracy percentage, class precision, and execution time for three algorithms (Random Forest, Naïve Bayes and Decision Tree) on the dataset after doing a detailed descriptive analysis of the data. The descriptive analysis was very insightful about the dataset. For example, Depari et al. (2022) found that the data contained mostly married females who like to travel alone on a sunny day around 6 PM. Most of them have attended college, yet didn’t graduate (Depari et al., 2022). For those who have an occupation, it states that most of them earn an income of around $25000 - $37499 (Depari et al., 2022). It was also mentioned that the destination is mostly the No Urgent Place such as Coffee House, which provides a coupon that expires in one day (Depari et al., 2022). It will be interesting to find more inferences from the data with the help of plots for each feature. This will help with finding the limitations of the dataset provided. Depari et al. (2022) found that predictive analytics results showed that random forest achieved the highest accuracy with 77.65% overall accuracy percentage, yet required the most time to process. However, the decision tree algorithm acquired the highest confidence level of 0.750 for prescriptive analysis (Depari et al., 2022). Prescriptive analytics is data analysis that is used to explain and determine what is the next action plan and why should they do it (Depari et al., 2022). It will be interesting to compare with 2 more algorithms: Logistic Regression and k-Nearest Neighbours (k-NN) used on the dataset. There was no dimension reduction done by Depari et al. (2022).

This dataset was also used by Atiq et al. (2022). This dataset comes with missing values. This could cause problems in the prediction analysis. To circumvent this problem Atiq et al. (2022) analysed the impact of four different imputation techniques (Frequent value, mean, KNN, MICE) to replace the missing values and use them to create prediction models. Atiq et al. (2022) then applied six classifier algorithms (Naïve Bayes, Deep Learning, Logistic Regression, Decision Tree, Random Forest, and Gradient Boosted Tree). Atiq et al. (2022) found that KNN imputation with Deep Learning classifier gave the most accurate outcome. Atiq et al. (2022) applied SMOTE (Synthetic Minority Oversampling Technique) for oversampling the dataset in order to have positive and negative instances divided into 50% / 50%. This led to a perfectly balanced dataset for target value. Below is a table by Atiq et al. (2022) that is used to show Accuracy of Classifiers on Actual and Oversampled Dataset using various imputation Techniques. **These pre-existing findings will be useful in dealing with the missing values in the dataset.** Also, there was no dimension reduction done by Atiq et al. (2022).

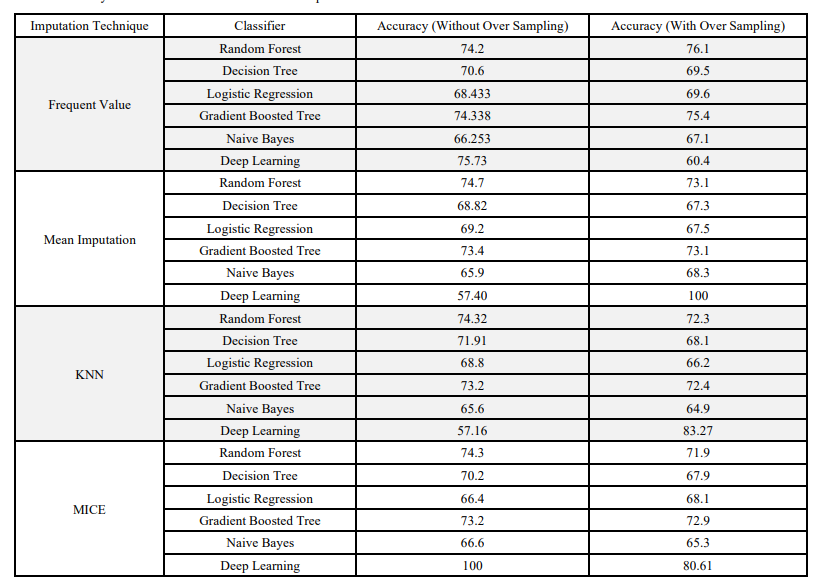


Table 1: Accuracy of Classifiers on Actual and Over Sampled Dataset provided by Atiq et al. (2022).

Patil et al. (2019) used a different dataset for E-coupons to predict coupon usage behaviour. It is interesting that to train the gradient boosting classifier, they used 45 “train periods” that simulated the test timing (Patil et al., 2019). When comparing with other algorithms like logistic regression, SVM, random forest, neural networks, Patil et al. (2019) found that gradient boosting was the single best classifier. By having train periods, Patil et al. (2019) made interesting observations such as the more the coupon is viewed, the probability to buy using the coupon code increases. Patil et al. (2019) also observed that customers tend to purchase the same coupon over and over again. The dataset for in-vehicle coupon response deficient in data over a periodic basis to help uncover such patterns. This is a limitation of the in-vehicle coupon dataset.

Ahmed et al. (2024) used a different dataset, Dunnhumby data for a particular grocery retail company. Ahmed et al. (2024) proposed two different models: one for predicting customer churn and the other for coupon redemption model and both those models used XGBoost Classifier Model. As seen in earlier papers, it is good to use a variety of models for comparison. However, Ahmed et al. used only XGBoost Classifier Model. Ahmed et al. (2024) also stated that their dataset was not large enough and therefore it led to uncertainty about the model performance on a larger dataset in the future.

Given all the existing data analysis that has been done on coupon redemption, it is obvious that it is crucial for businesses to help them grow and survive. When we attain a dataset with fewer dimensions using these 3 methods: Stepwise Regression, Forward Feature Selection and Backward Feature Elimination methods learnt by us in *CMTH 642 – Data Analytics: Advanced Methods*, we can have more concise information to be used for decision making and not have noise. **Dimensionality reduction is the major contribution of this project to this dataset.** A visual correlation matrix will show which attributes are highly correlated to the target of the customer accepting or rejecting a coupon. A visual correlation matrix is a very effective way to display a lot of information. This project also gives an opportunity to compare and find the best predictive classification algorithm for In-Vehicle Coupon Recommendation dataset (2020) after evaluation of various supervised learning classification algorithms introduced to us in *CMTH 642 – Data Analytics: Advanced Methods* like Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-Nearest Neighbours (k-NN) on the dataset. There could be so many other factors that effect customer coupon redemption. Therefore, we need to also explore the limitations of this dataset. Hence, it is worth proceeding with this project.

# Data Description

The dataset is in csv format. It is a publicly available dataset called In-Vehicle Coupon Recommendation at UCI Machine Learning Repository from 2020 which describes different driving scenarios of multiple clients and whether the coupon is accepted (*UCI Machine Learning Repository*, 2020). The dataset is large with 12684 records and 25 features. The below figure provides some information on the dataset:

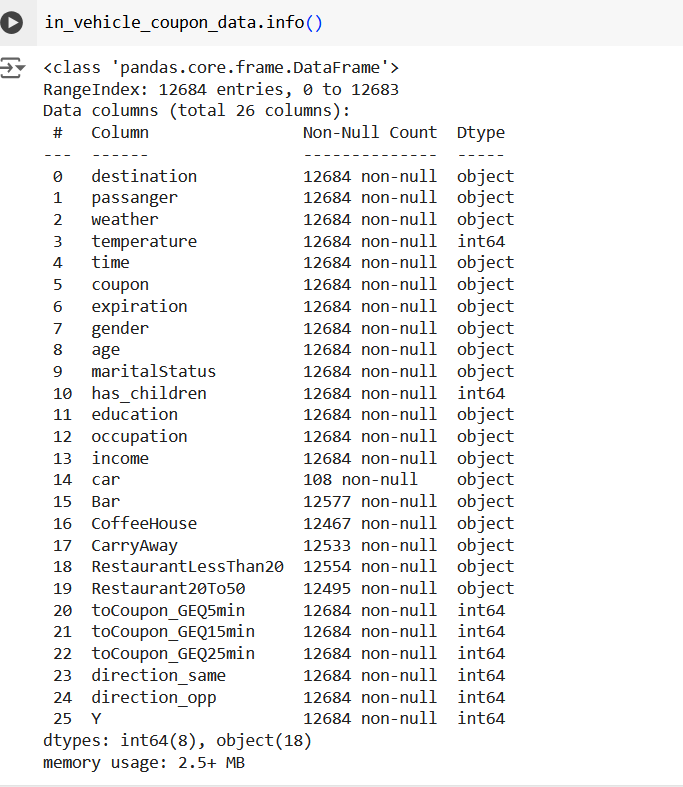


Figure 4: In-Vehicle Coupon Recommendation dataset information

|  |  |  |
| --- | --- | --- |
| Feature | Feature Description | Possible Values |
| destination | Destination | No Urgent Place, Home, Work |
| passenger | Passenger | Alone, Friend(s), Kid(s), Partner |
| weather | Weather Type | Sunny, Rainy, Snowy |
| temperature | Temperature | Numerical value |
| time | Time | 2PM, 10AM, 6PM, 7AM, 10PM |
| coupon | Coupon type | Restaurant (<20), Coffee House, Carry out & Take away, Bar, Restaurant (20-50) |
| expiration | Expiration | 1d, 2h |
| gender | Gender | Female, Male |
| age | Age | below21, 50plus, 36, 41, 31, 26, 46, 21 |
| maritalStatus | Marital Status | Unmarried partner, Single, Married partner, Divorced, Widowed |
| has\_children | Whether they have children | Numerical value – 0 or 1 |
| education | Education | Some college – no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), some High School |
| occupation | Occupation | Unemployed, Architecture & Engineering, Student, Education & Training & Library, Healthcare Support, Healthcare Practitioners & Technical, Sales & Related, Management, Arts Design Entertainment Sports & Media, Computer & Mathematical, Life Physical Social Science, Personal Care & Service, Community & Social Services, Office & Administrative Support, Construction & Extraction, Legal, Retired, Installation Maintenance & Repair, Transportation & Material Moving, Business & Financial, Protective Service, Food Preparation & Serving Related, Production Occupations, Building & Grounds Cleaning & Maintenance, Farming Fishing & Forestry |
| income | Income | 37500-49999, 62500-74999, 12500 – 24999, 75000-87499, 50000 – 62499, 25000 – 37499, $100000 or More, 87500 – 99999, Less than $12500 |
| car | Car | Scooter and motorcycle, crossover, Mazda5, do not drive, car that is too old to install Onstar |
| bar | The number of times they go to the bar per month | Never, less 1, 1~3, gt8, 4~8 |
| CoffeeHouse | The number of times they go to a coffee shop per month | Never, less1, 4~8, 1~3, gt8 |
| CarryAway | The number of times that they buy take away food per month | Never, less1, 4~8, 1~3, gt8 |
| RestaurantLessThan20 | if customer’s average expense per person at restaurants is less than 20 dollars a month | Never, less1, 4~8, 1~3, gt8 |
| Restaurant20To50 | if customer’s average expense per person at a restaurant is between 20 dollars to 50 dollars per month | Never, less1, 4~8, 1~3, gt8 |
| toCoupon\_GEQ5min | driving distance to the restaurant/bar for using the coupon is greater than 5 minutes | Numerical value |
| toCoupon\_GEQ15min | driving distance to the restaurant/bar for using the coupon is greater than 15 minutes | Numerical value |
| toCoupon\_GEQ25min | driving distance to the restaurant/bar for using the coupon is greater than 25 minutes | Numerical value |
| direction\_same | whether the restaurant/bar is in the same direction as destination | Numerical value |
| direction\_opp | whether the restaurant/bar is in the opposite direction as destination, | Numerical value |
| Y | whether the coupon is accepted. This is the target. | 0 or 1. 1 if coupon is accepted. 0 if coupon is rejected. |

Table 2: Feature Description in In Vehicle Coupon Redemption dataset

The below plots for distribution of categorical value features were generated with the help of code by Inyama (2023).

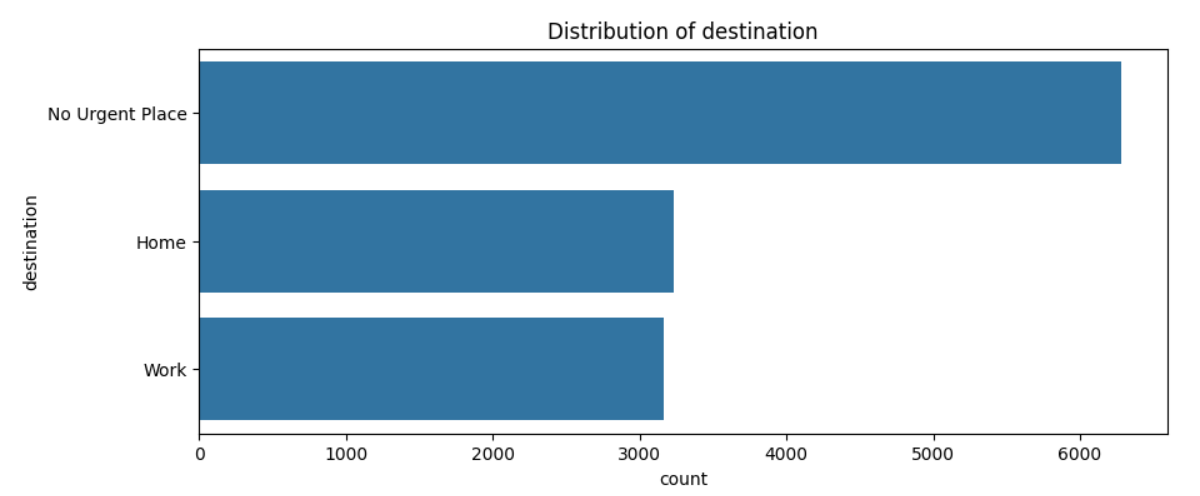


Figure 5: Distribution of destination feature

Observation: Most people’s destination is not an urgent place

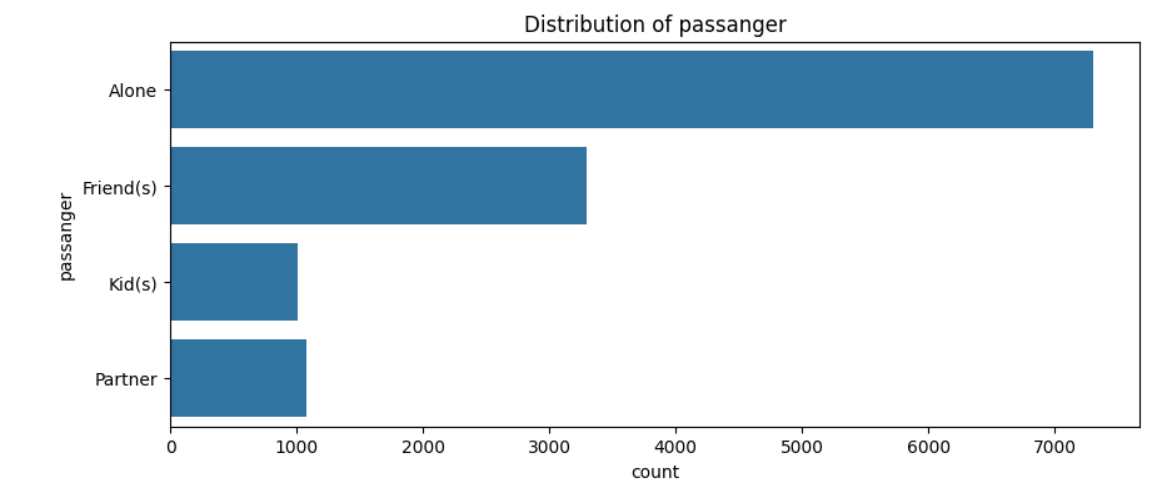


Figure 6: Distribution of passenger feature

Observation: Most people travel alone

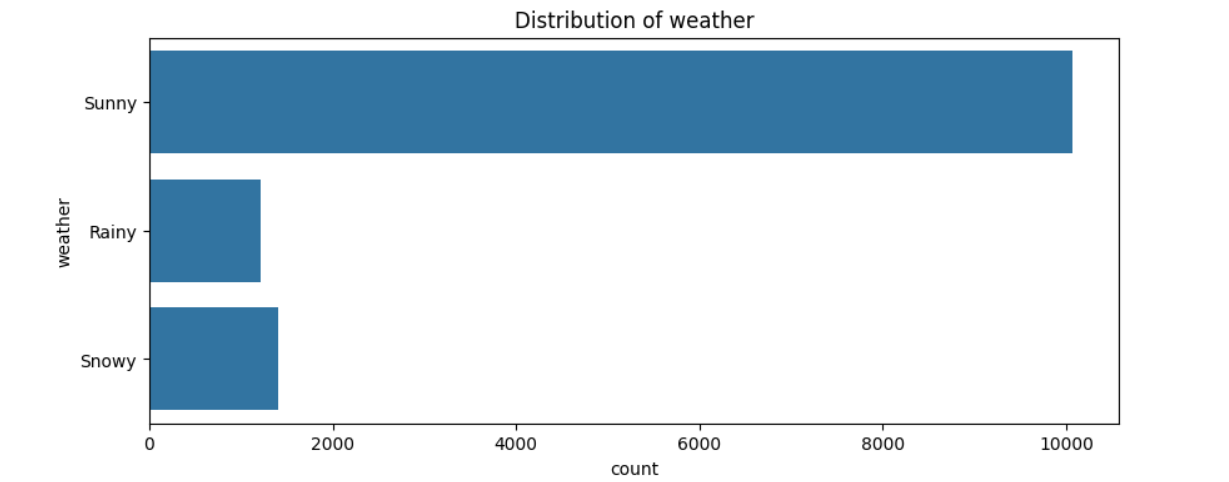


Figure 7: Distribution of weather feature

Observation: Most people travel on a sunny day

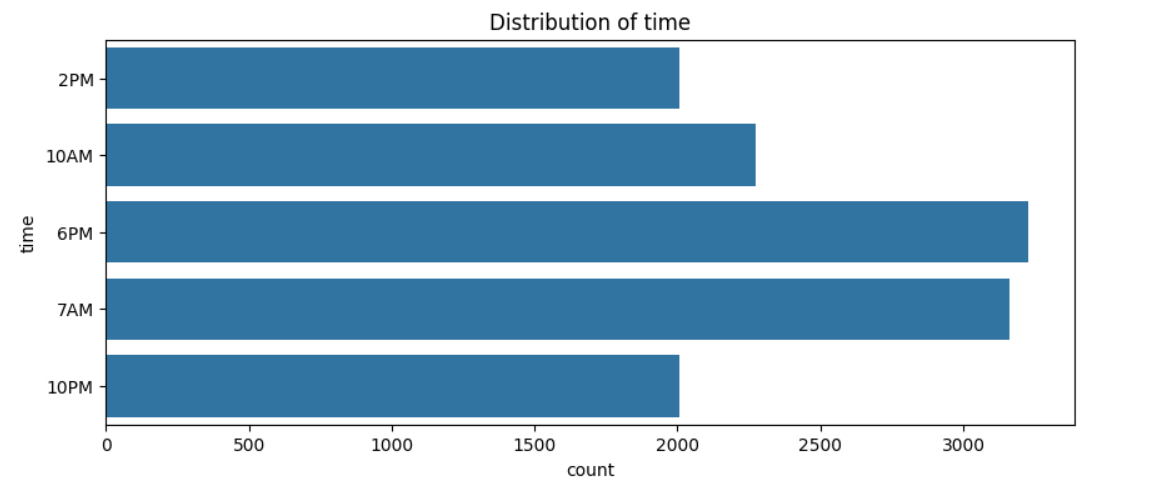


Figure 8: Distribution of time feature

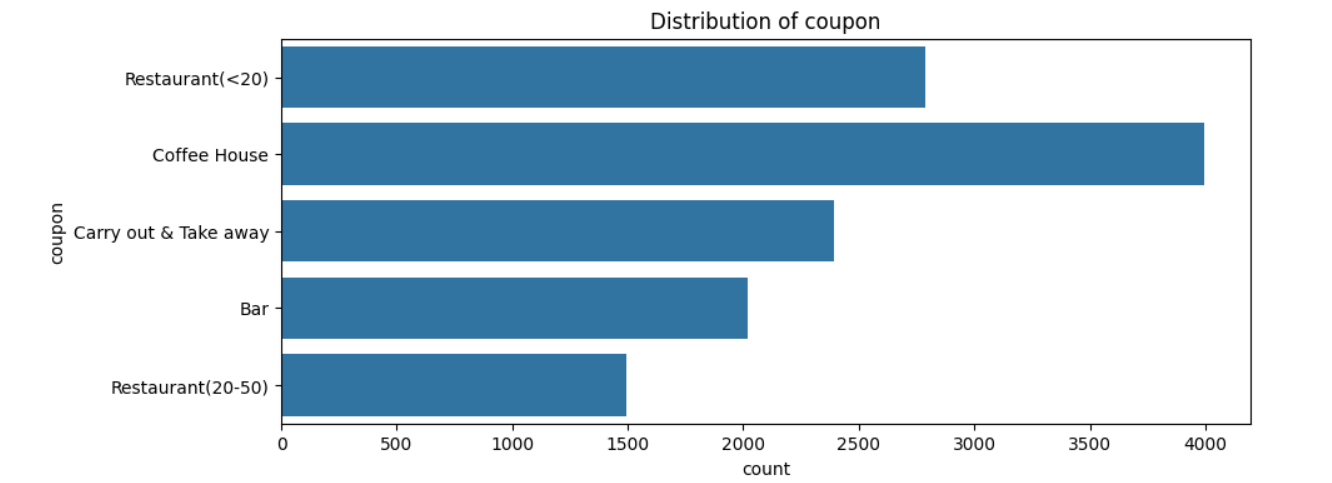


Figure 9: Distribution of coupon feature

Observation: Most people go to Coffee House

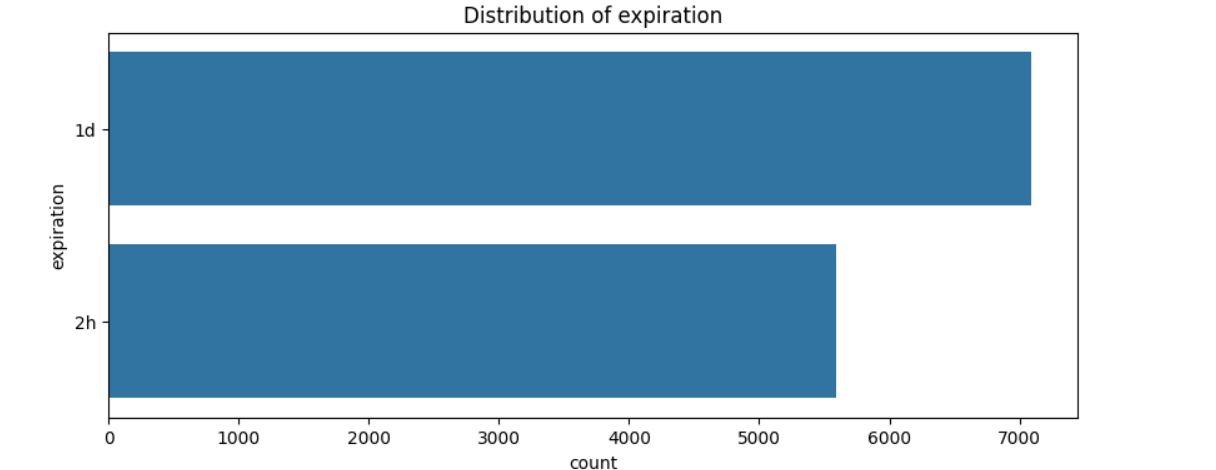


Figure 10: Distribution of expiration feature

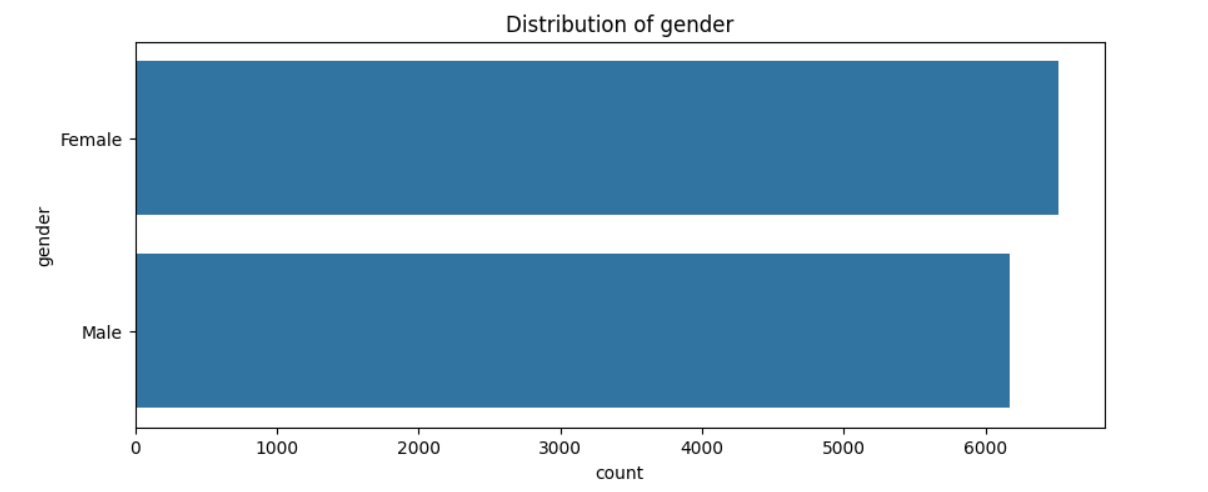


Figure 11: Distribution of gender feature

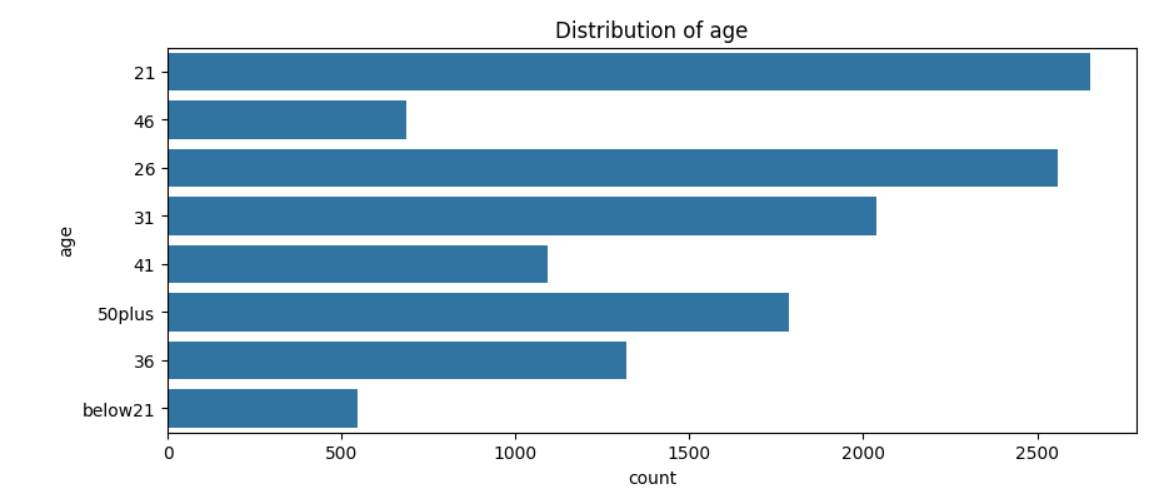


Figure 12: Distribution of age feature

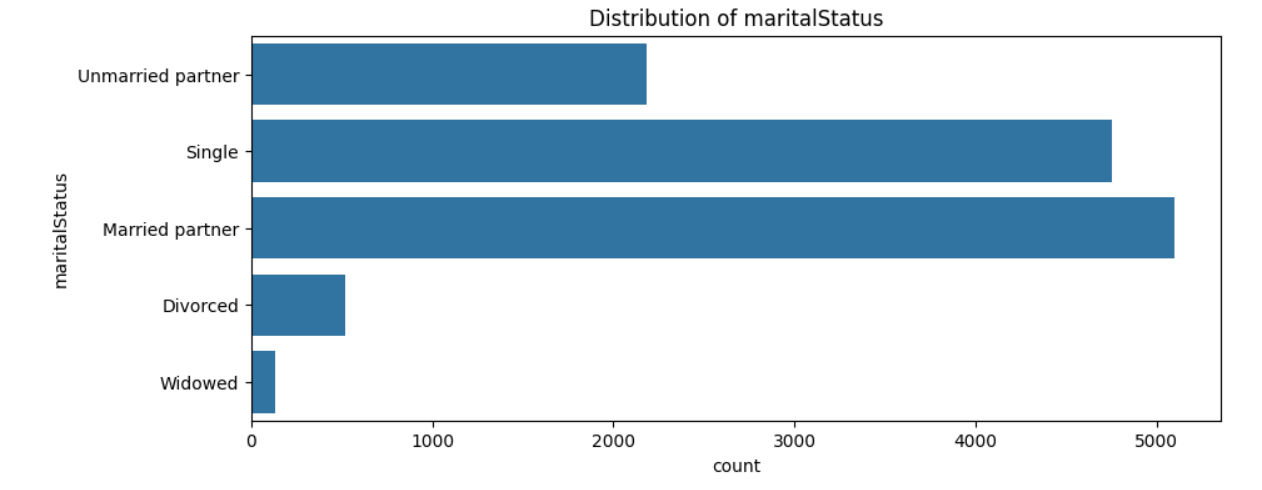


Figure 13: Distribution of marital status feature

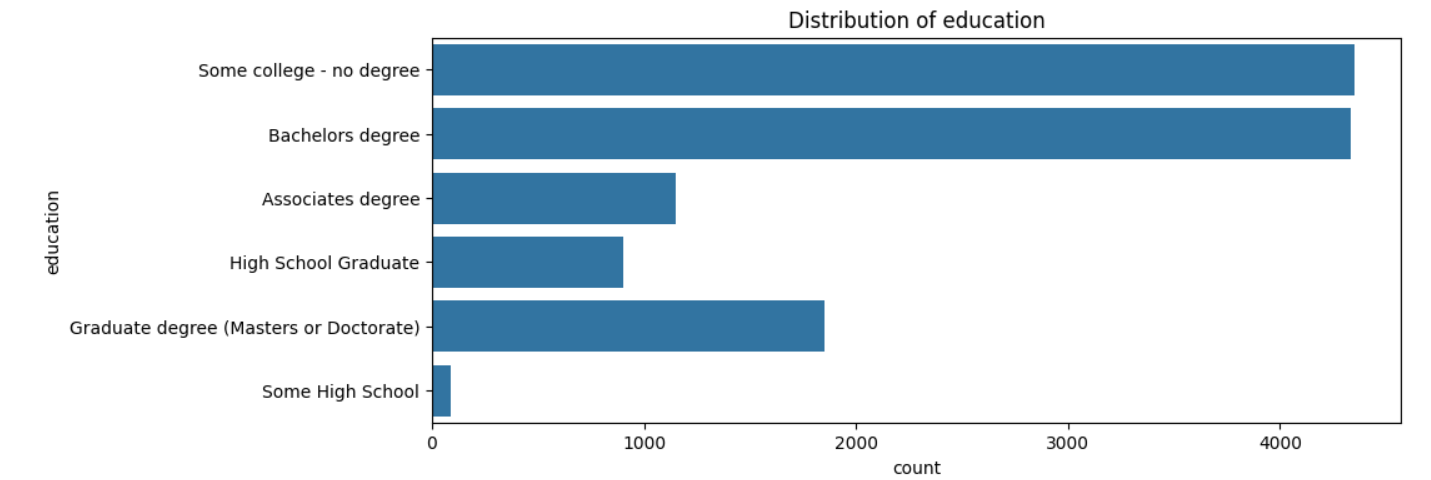


Figure 14: Distribution of education feature

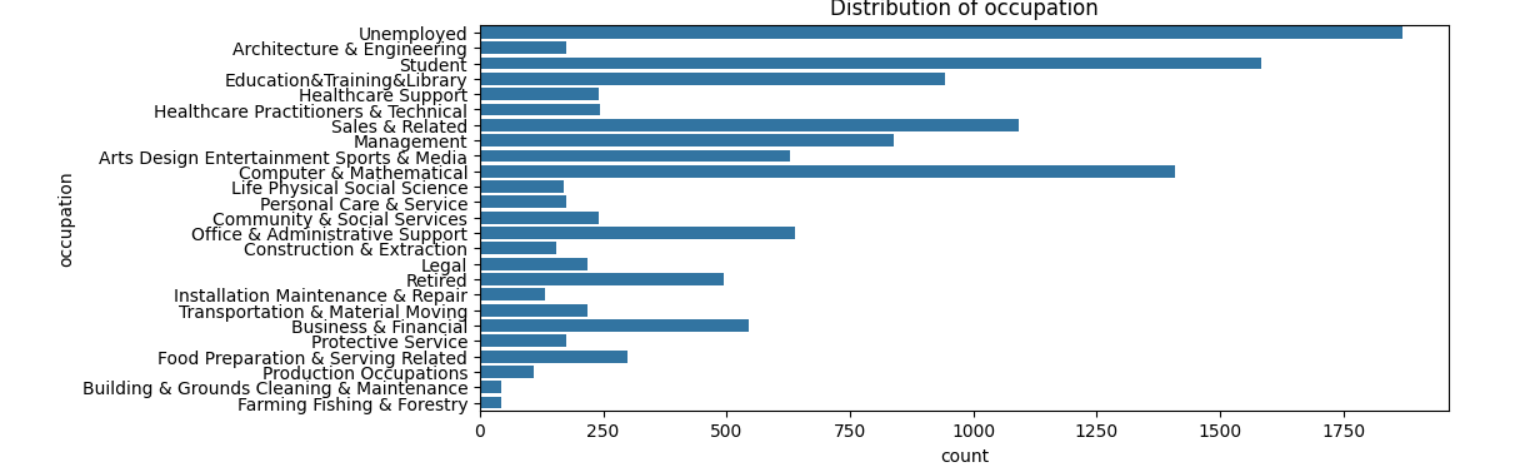


Figure 15: Distribution of occupation feature

Observation: Occupation has too many categories.

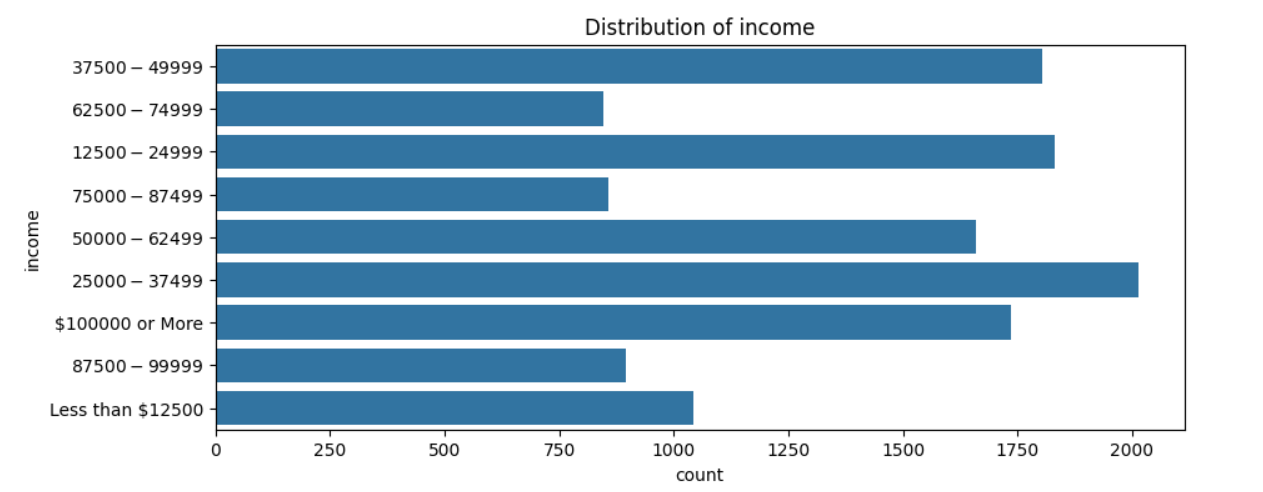


Figure 16: Distribution of income feature

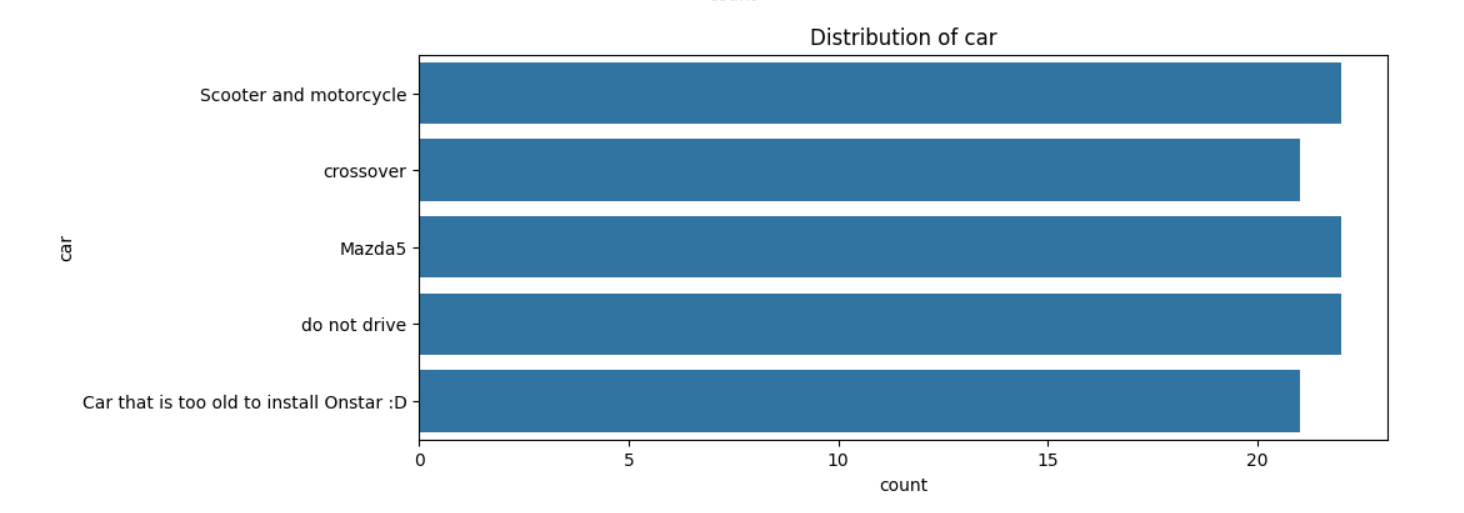


Figure 17: Distribution of car feature

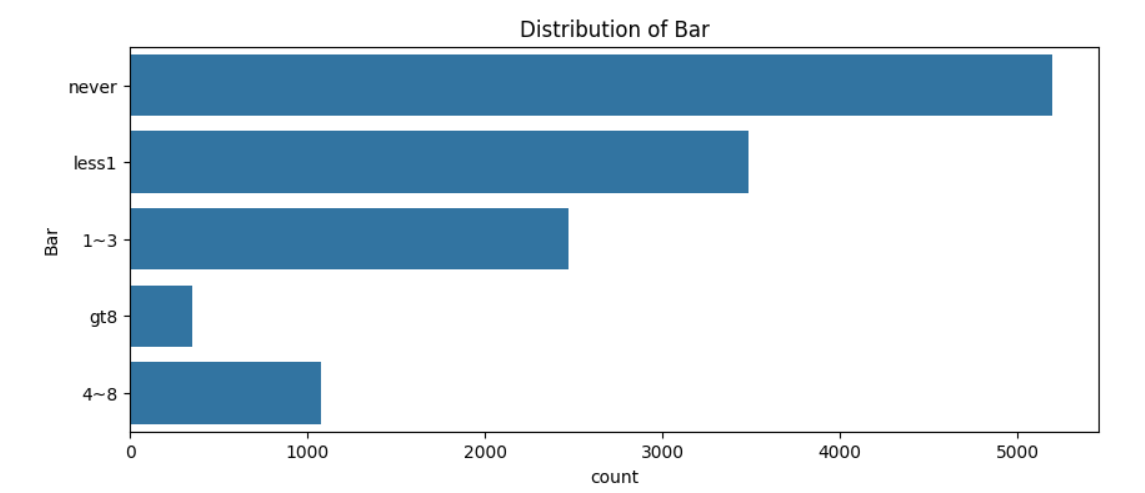


Figure 18: Distribution of bar feature

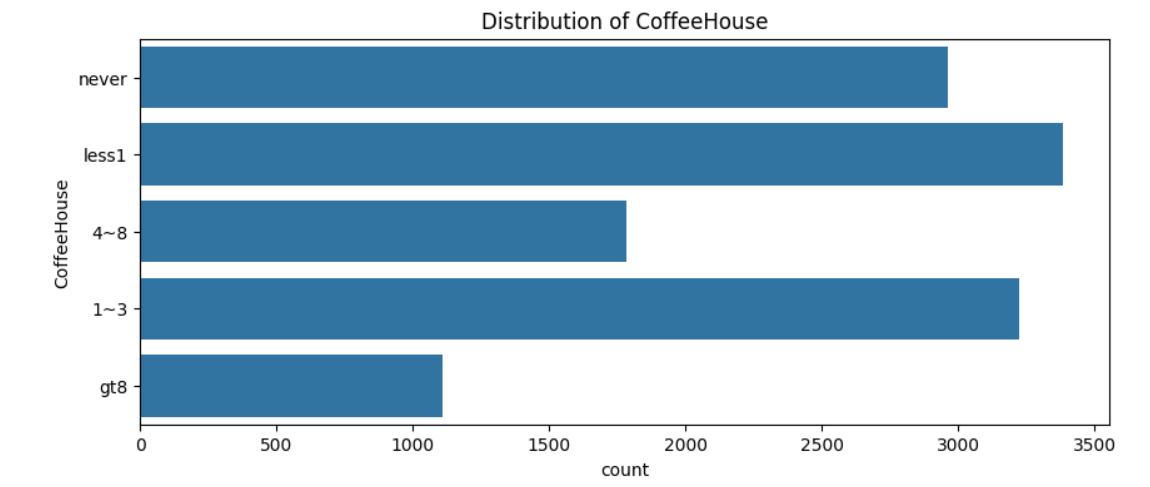


Figure 19: Distribution of Coffee House feature

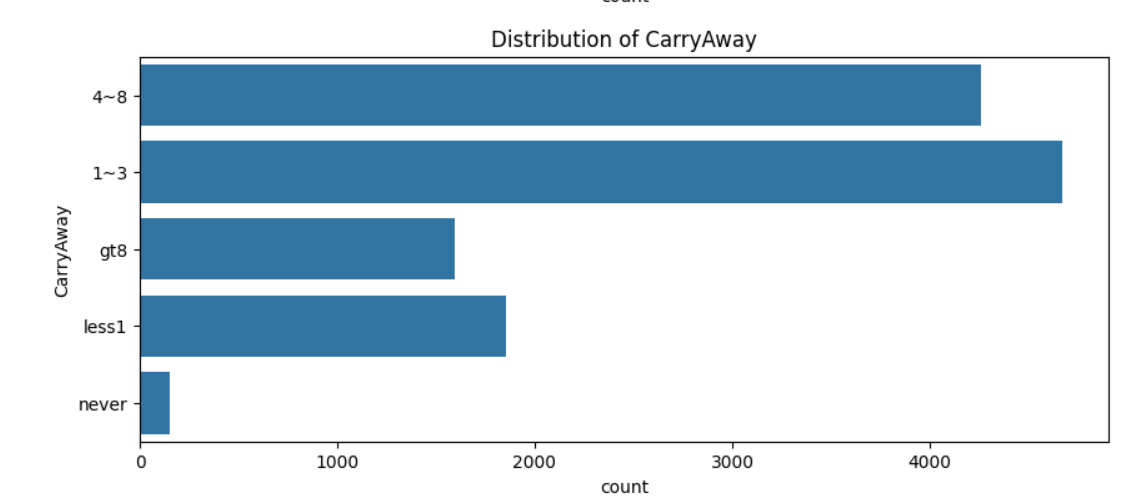


Figure 20: Distribution of Carry Away feature

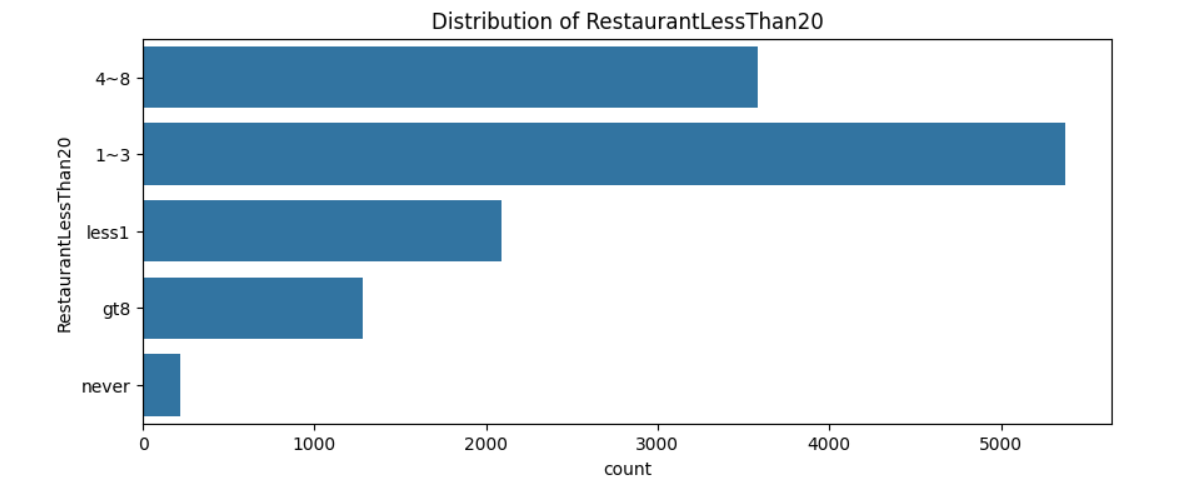


Figure 21: Distribution of Restaurant Less Than 20 feature

Observation: Most popular frequency for RestaurantLessThan20 feature is 1 to 3

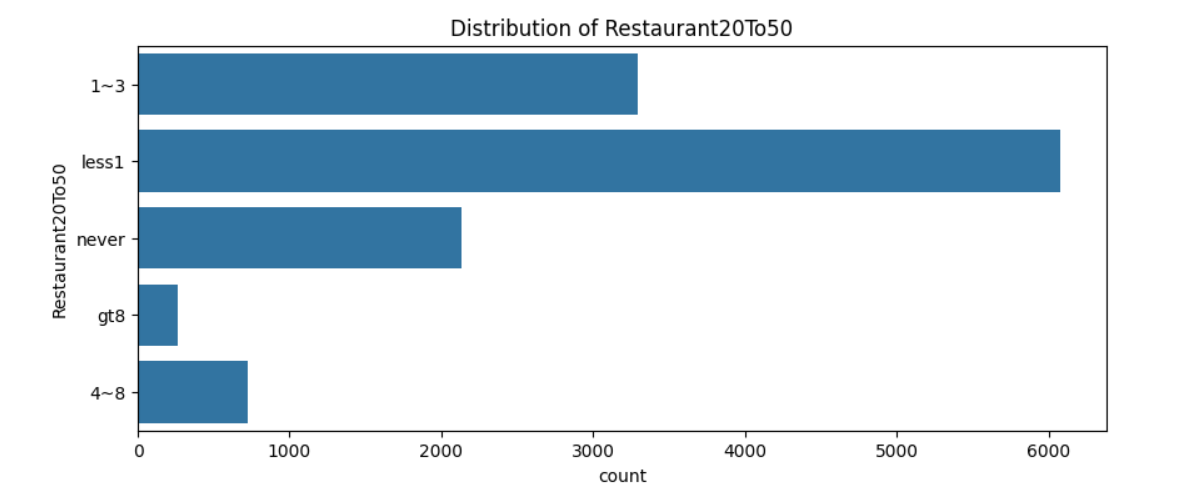


Figure 22: Distribution of Restaurant 20 to 50 feature

Observation: Most popular frequency for Restaurant20To50 feature is less 1

Using code from Inyama (2023), we can see the boxplots of the numerical data in numerical columns.

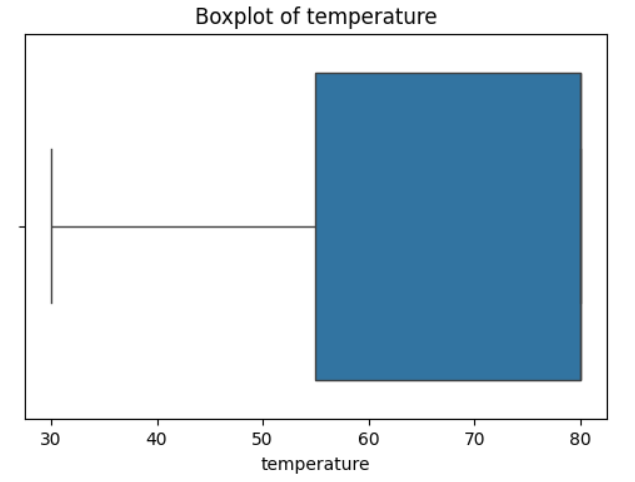


Figure 23: Boxplot of temperature

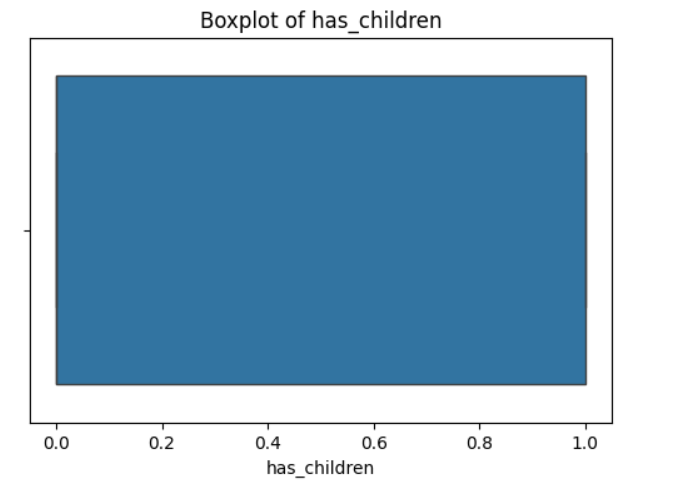


Figure 24: Boxplot of has children

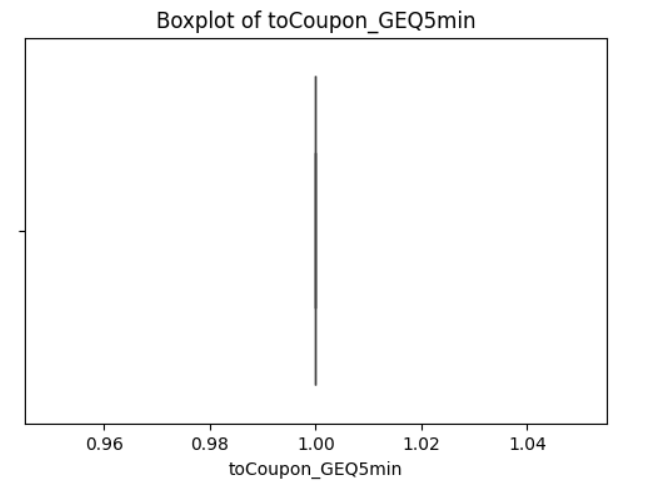


Figure 25: Boxplot of to coupon GEQ 5 min

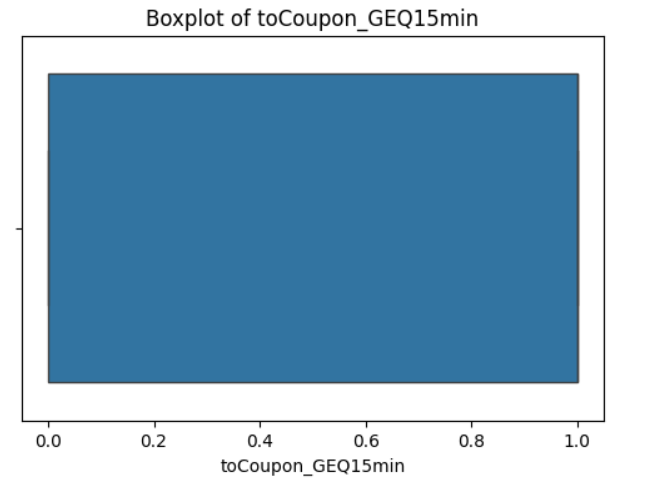


Figure 26: Boxplot of to coupon GEQ 15 min

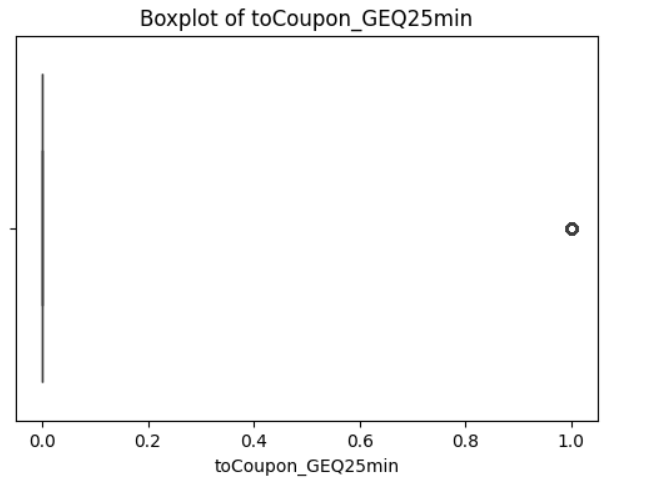


Figure 27: Boxplot of to coupon GEQ 25 min

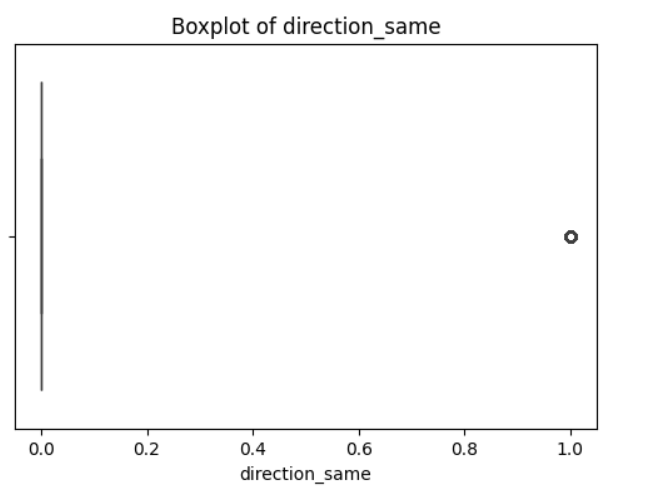


Figure 28: Boxplot of direction same

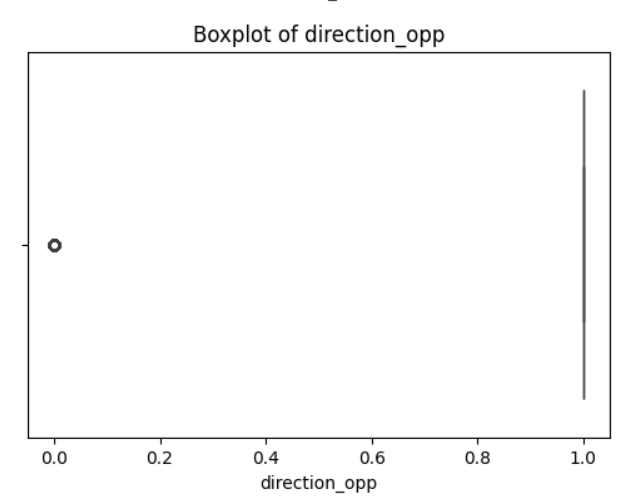


Figure 29: Boxplot of direction opposite

Using code provided by Niralidedaniya (2023), it was found that the target classes are partially balanced. If the target classes were highly unbalanced, then this dataset could not be used because the results of supervised learning algorithms used to make predictions would skew towards the class with the class with higher percentage of records.



Figure 30: Distribution of Target Classes

It should also be noted that this dataset comes with many missing values. Hence, this dataset requires preprocessing before it can be analyzed with machine learning algorithms. Using the code provided by Niralidedaniya (2023), we can see the features which have missing values.

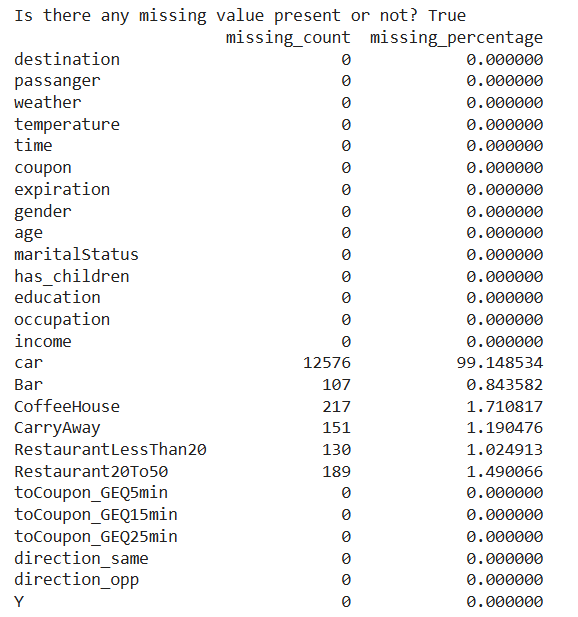


Figure 31: Distribution of missing values in the dataset

There are too many missing values for car feature. To avoid unreliable results, the car feature will be dropped from the dataset. toCoupon\_GEQ5min has the same value for all rows. This does not add any useful information to the analysis. toCoupon\_GEQ5min will be dropped from the dataset. Occupation feature has too many categories. This creates too much noise and distracts from clear analysis. Occupation feature will be dropped from the dataset.

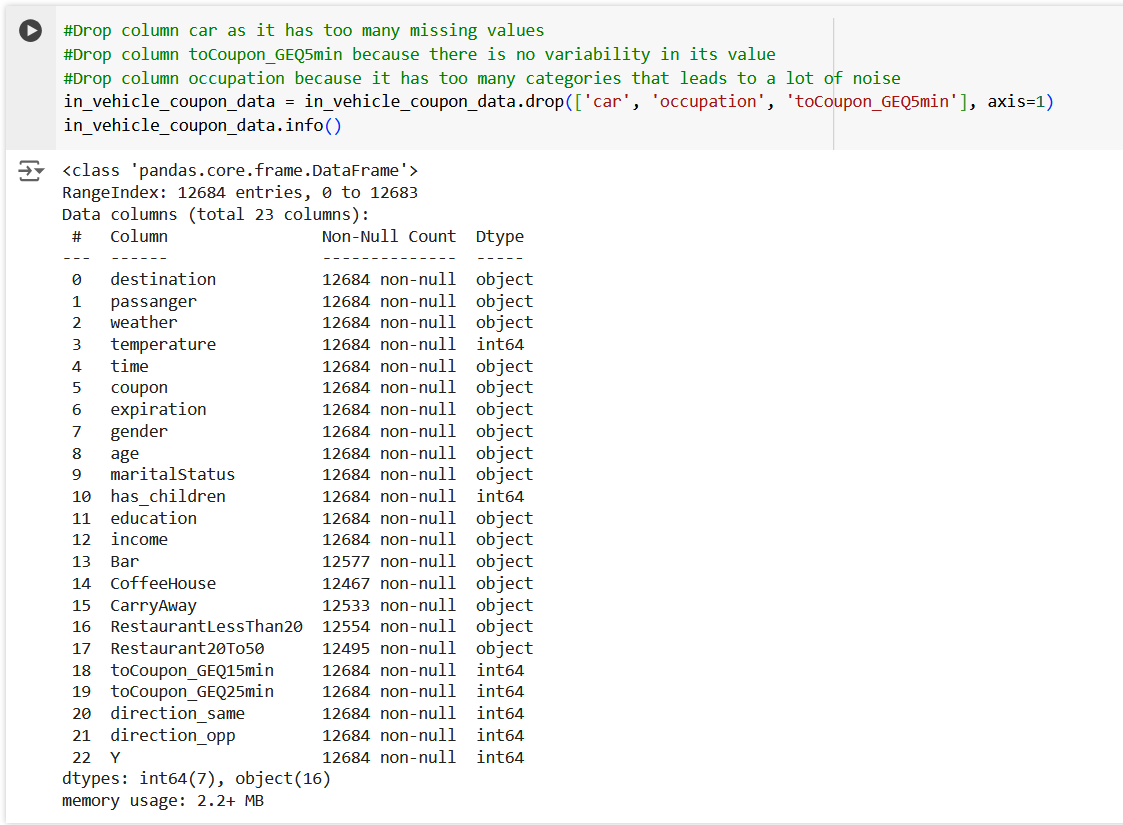


Figure 32: Information of dataset after dropping features: car, toCoupon\_GEQ5min, occupation

**From Table 1, Atiq et al. (2022) demonstrated pretty good accuracy when using frequent value imputation for missing values in the dataset along with the algorithms: Random Forest, Decision Tree, Logistic Regression, Gradient Boosted Tree, Naïve Bayes and Deep Learning.** Therefore, this project is going to use frequent value imputation too for the remaining missing values. This can be easily done with the help of code snippet provided by Niralidedaniya (2022).

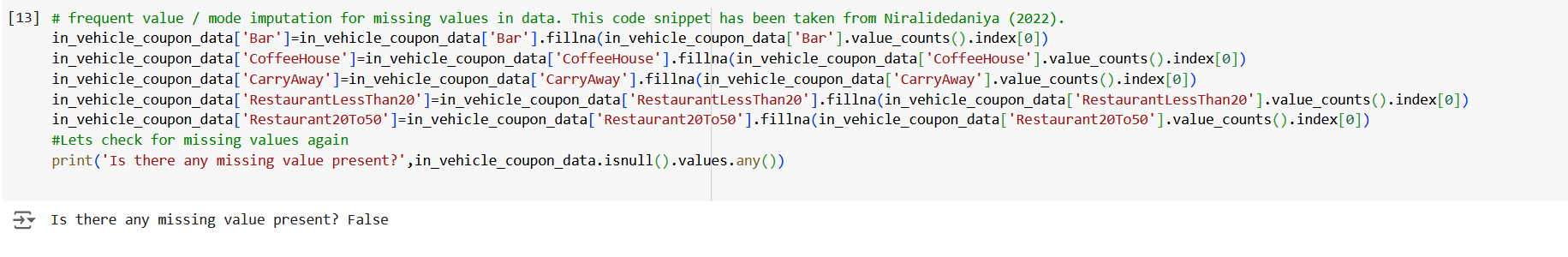
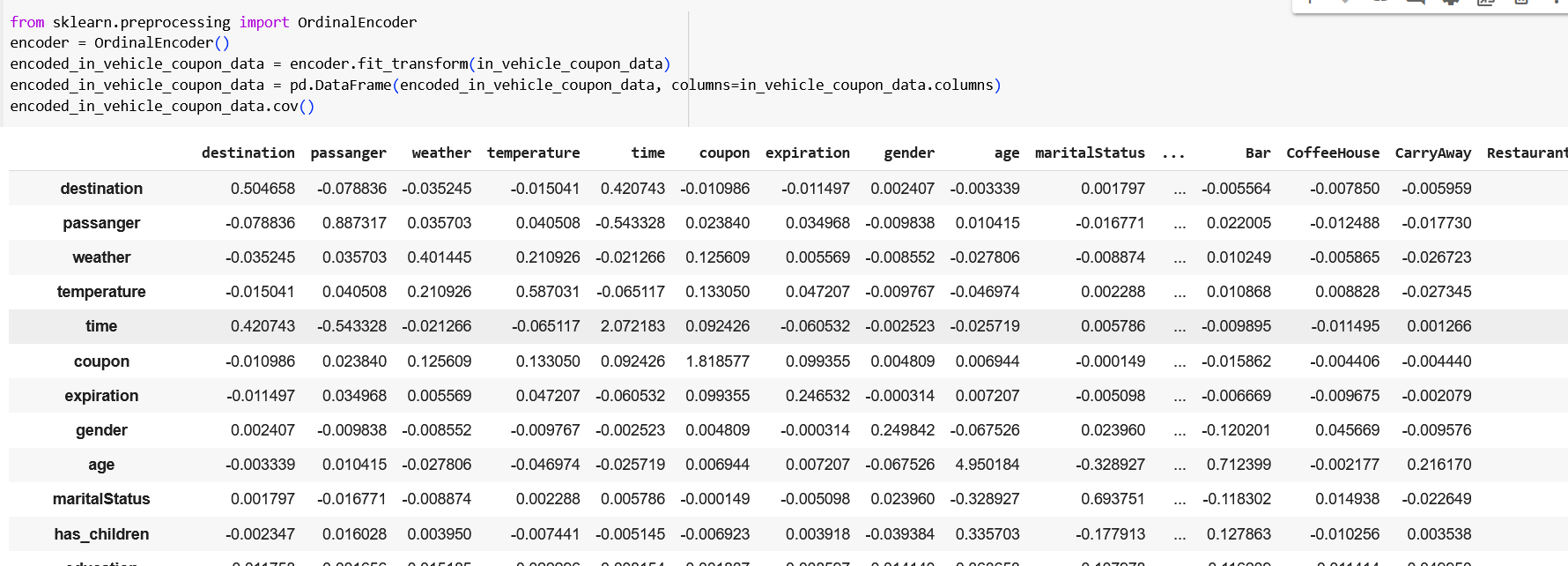


Figure 33: Mode / frequent value imputation for missing values

Using ordinal encoding for the categorical values in the categorical features, the covariance matrix was generated. Strangely, direction\_same feature has same values as direction\_opp feature in the covariance matrix.



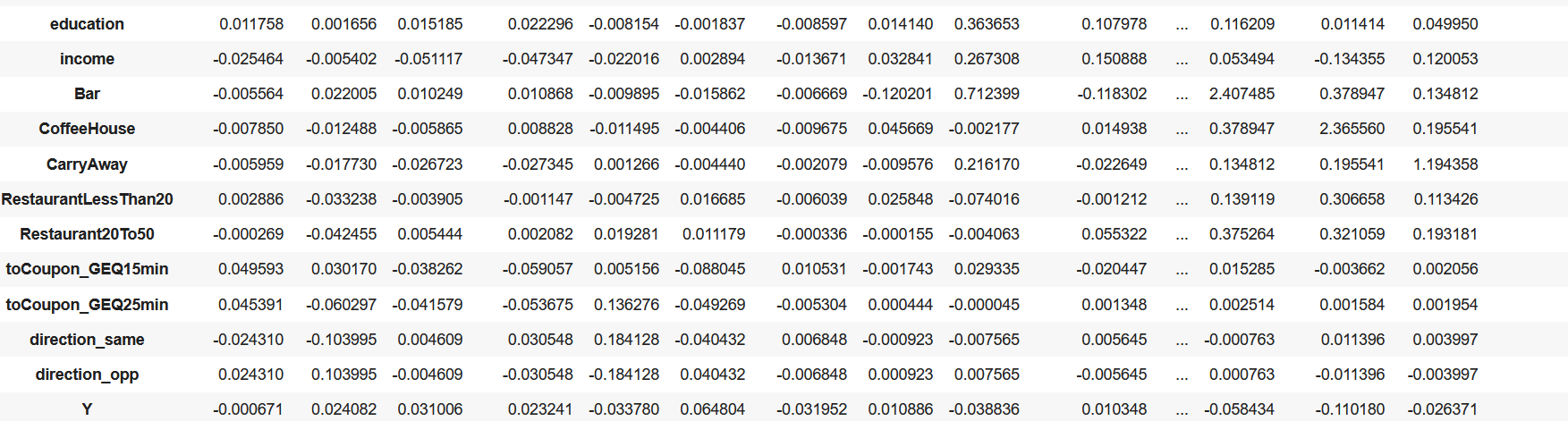


Figure 34: Covariance matrix on the dataset after doing ordinal encoding on the categorical values of the dataset.

Since direction\_same has same covariance values as direction\_opp. It makes sense to just have one on them and reduce the noise.

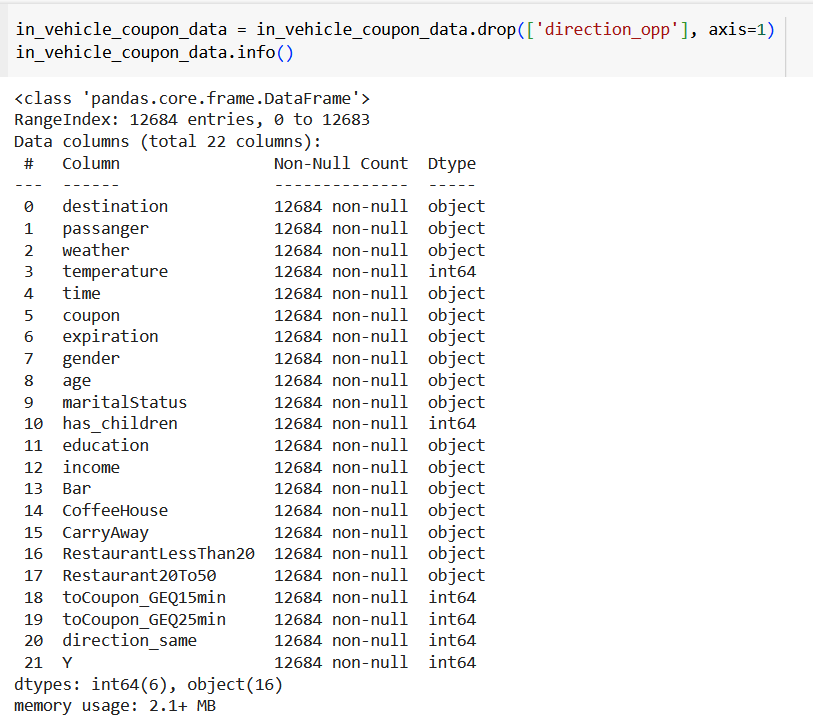
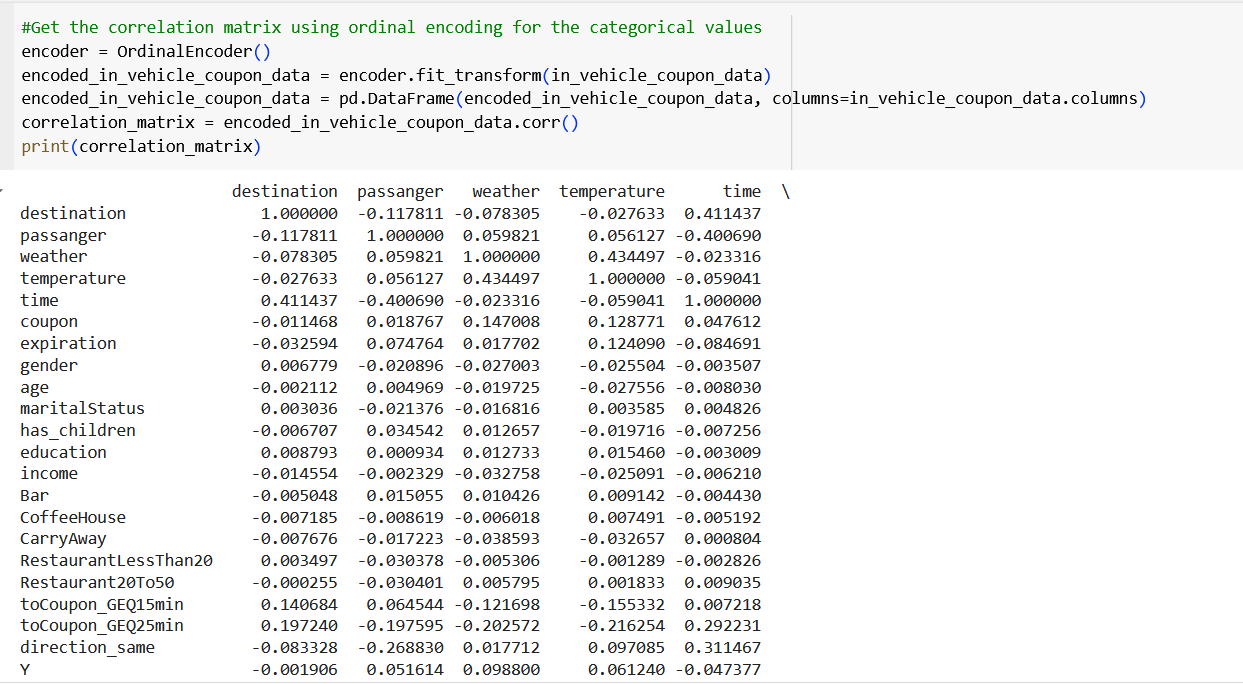
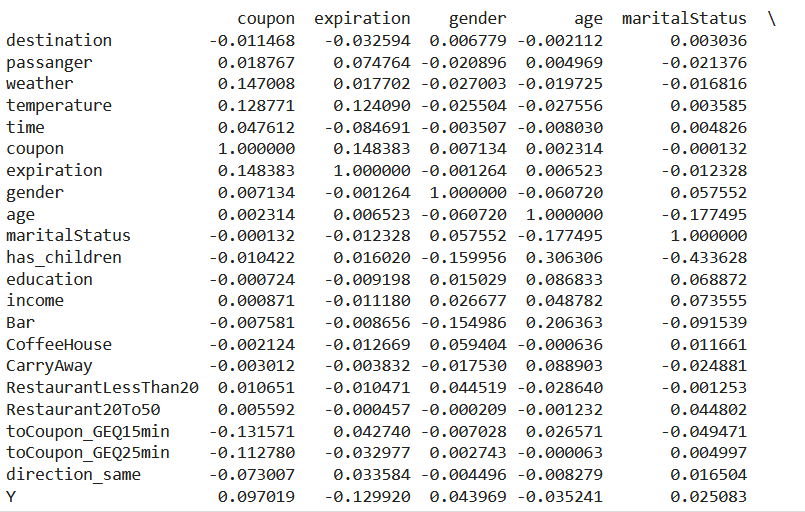
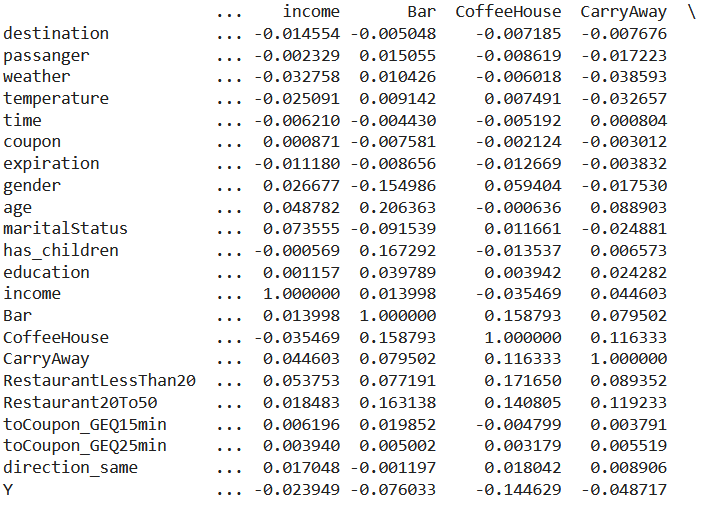


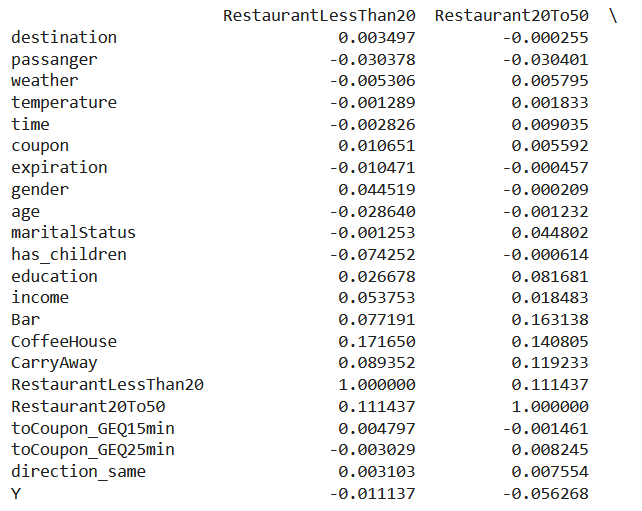
Figure 35: Dataset information after data cleaning

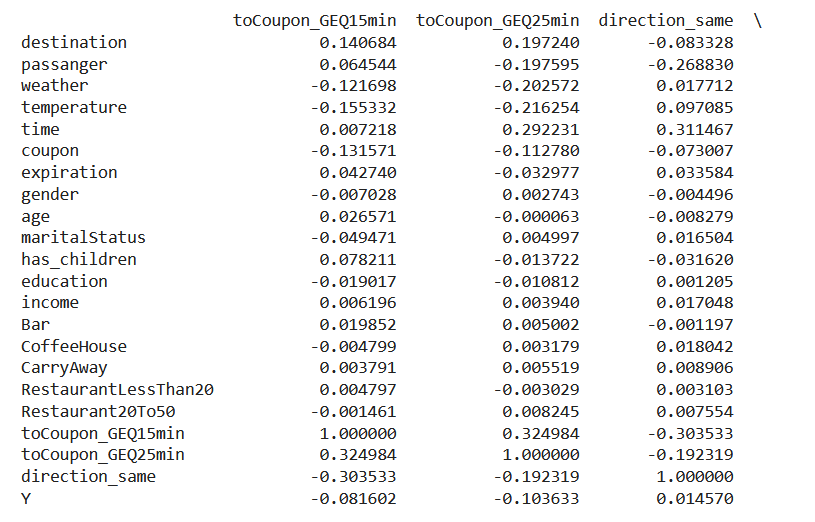
With all the steps above, the data is cleaned and ready to use. We can also generate a correlation matrix.











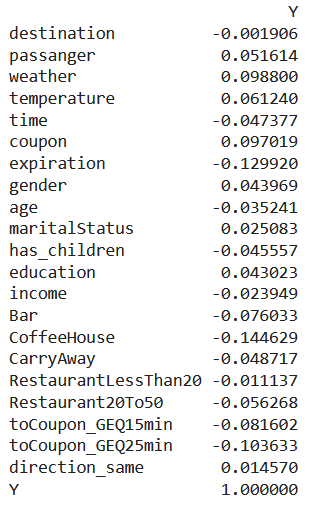


Figure 36: Correlation matrix after cleaning In Vehicle Coupon dataset

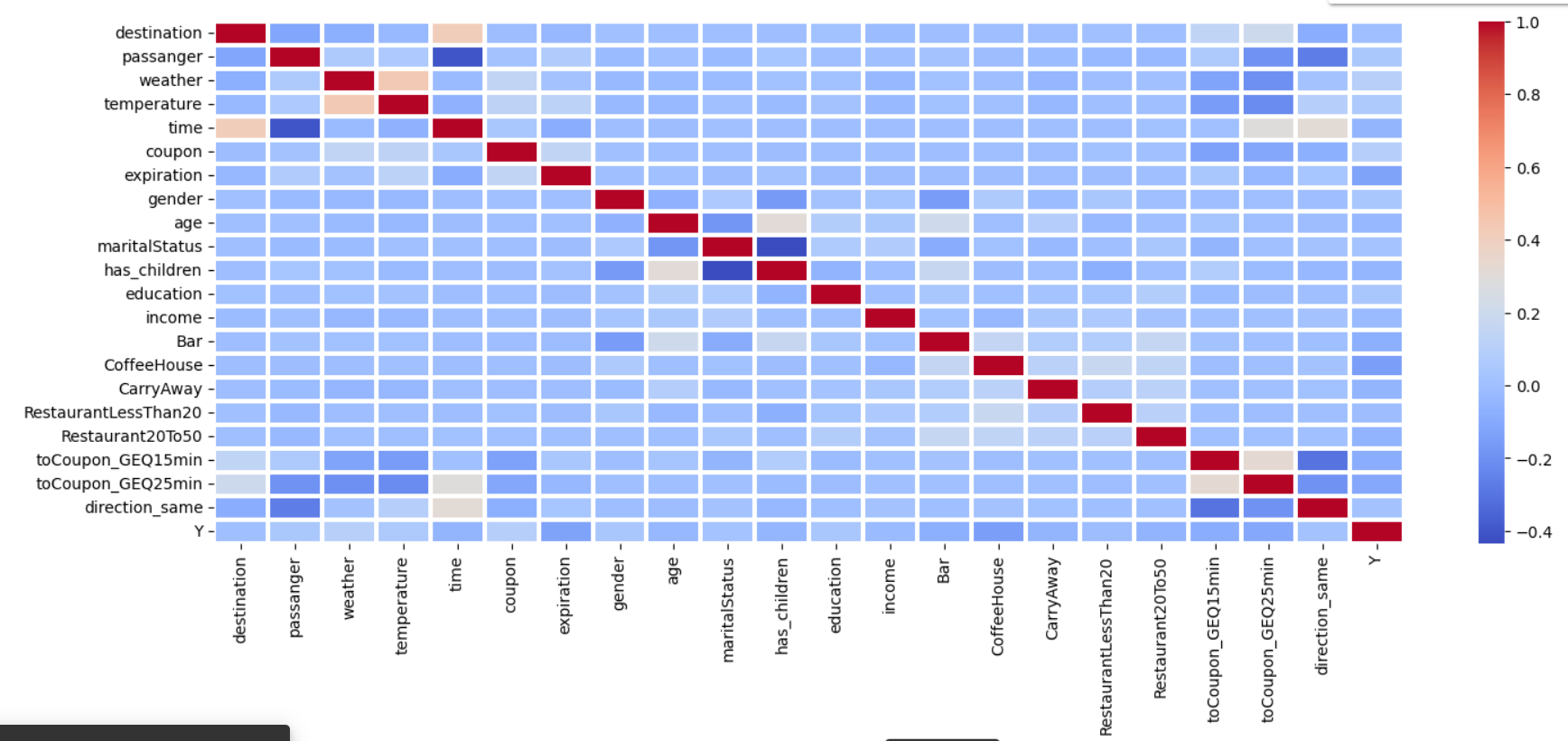


Figure 37: Heat map of correlation matrix

A heatmap to visualize the correlation between features is also effective.

There seems to be a correlation between time and destination, between temperature and weather, between marital status and has children, between passenger and time, between to coupon GEQ 15 min and same direction. **There seems to be correlation of the following features with the target Y which determines whether the coupon was accepted or rejected: expiration, CoffeeHouse, toCoupon\_GEQ15min, toCoupon\_GEQ25min. In the heat map, the darker the color, the stronger the correlation. This also corresponds with the higher values in the correlation matrix.**

# Dimensionality Reduction

Dimensionality reduction was done using 3 methods: Stepwise Regression, Forward Feature Selection and Backward Feature Elimination methods learnt by us in *CMTH 642 – Data Analytics: Advanced Methods*. This will help eliminate some noise. **Dimensionality reduction is the major contribution of this project to this dataset.** Past work as seen in the Literature Review did not do dimensionality reduction on this dataset.

**Stepwise Regression**

Alpha of 0.05 was chosen. Stepwise Regression for dimensionality reduction was done using Python. Coefficients having a p-value of 0.05 or less will be statistically significant.

**Iteration 1:** Bar has highest p-value of 0.779. Bar will be dropped. It is the least statistically significant.

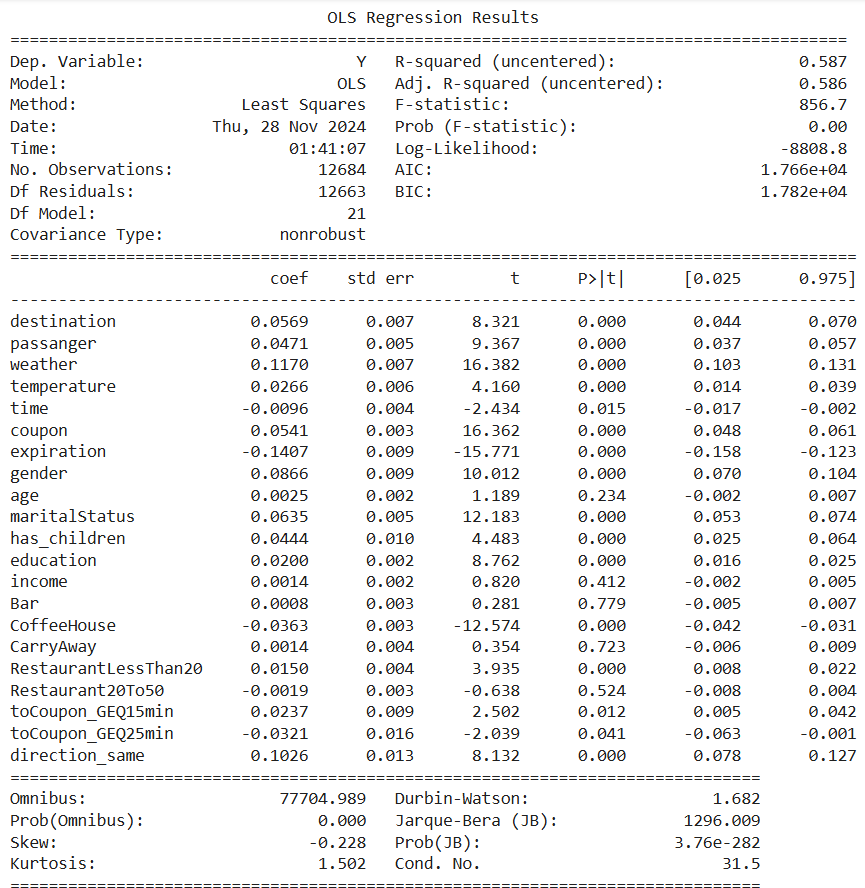


Figure 38: First iteration of stepwise regression

**Iteration 2:** CarryAway is least statistically significant because it has highest p-value of 0.713 which is greater than alpha of 0.05. CarryAway should be dropped.

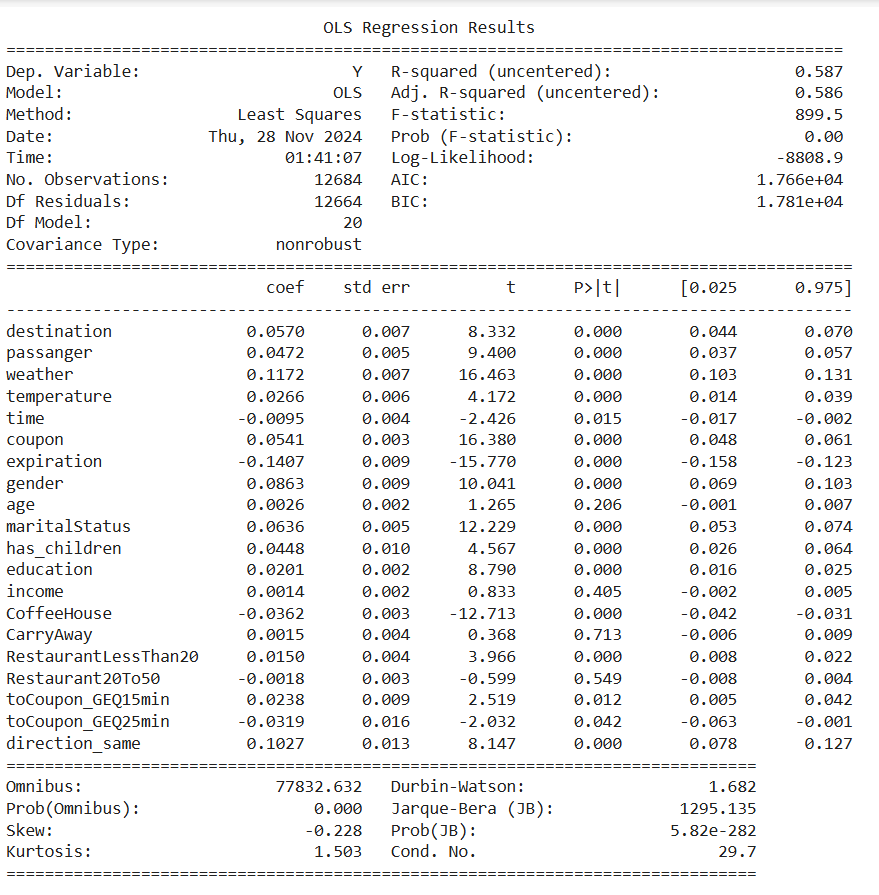


Figure 39: Second iteration of stepwise regression

**Iteration 3:** Restaurant20To50 is least statistically significant because it has highest p-value of 0.576 which is greater than alpha of 0.05. Restaurant20To50 should be dropped.

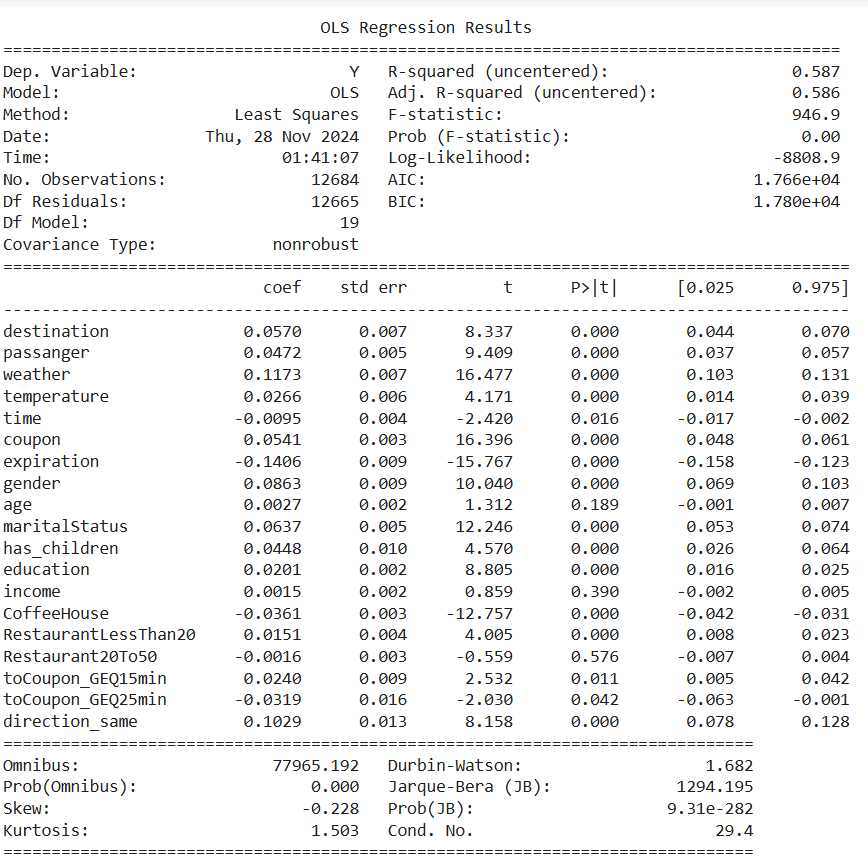


Figure 40: Third iteration of stepwise regression

**Iteration 4:** Income is least statistically significant because it has highest p-value of 0.405 which is greater than alpha of 0.05. Income should be dropped.

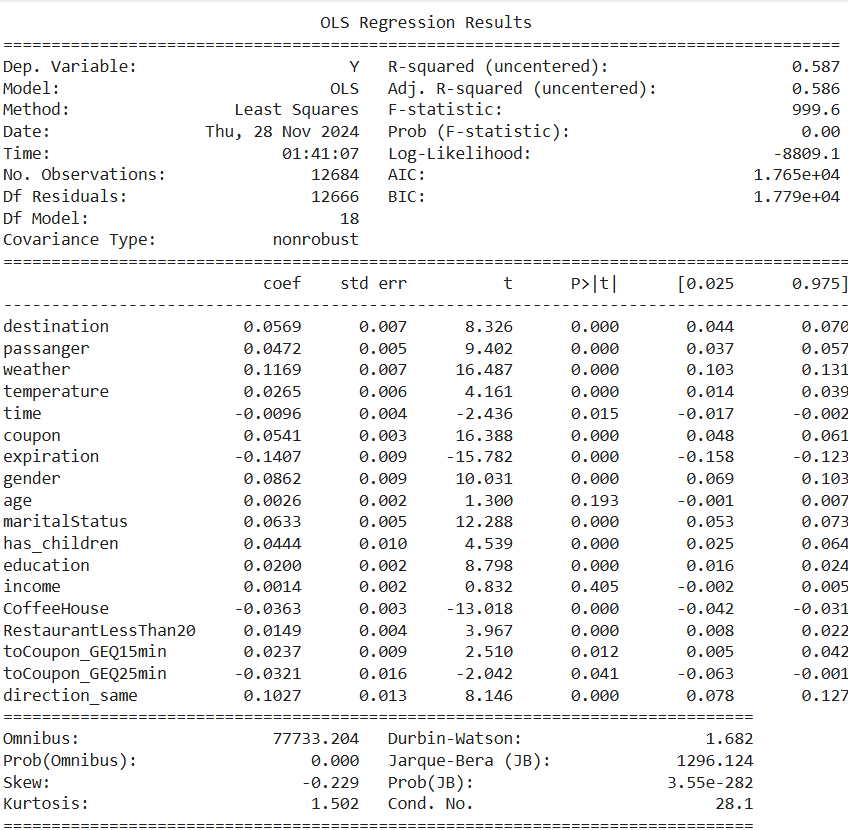


Figure 41: Fourth iteration of stepwise regression

**Iteration 5:** Age is least statistically significant because it has highest p-value of 0.166 which is greater than alpha of 0.05. age should be dropped.

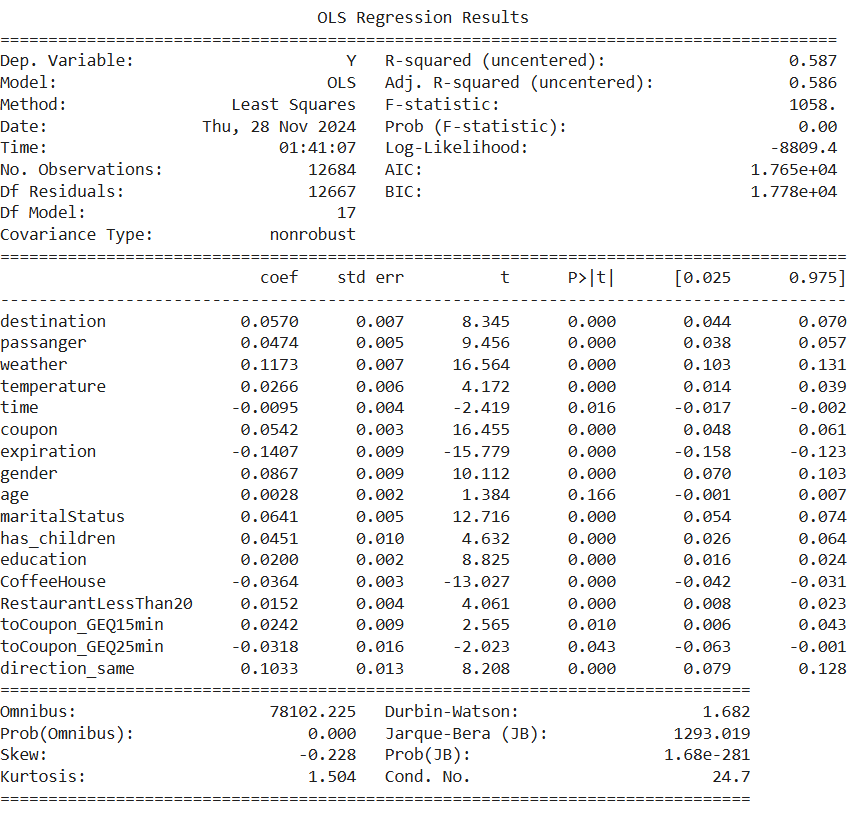


Figure 42: Fifth iteration of stepwise regression

After 5 iterations, all coefficients are statistically significant.

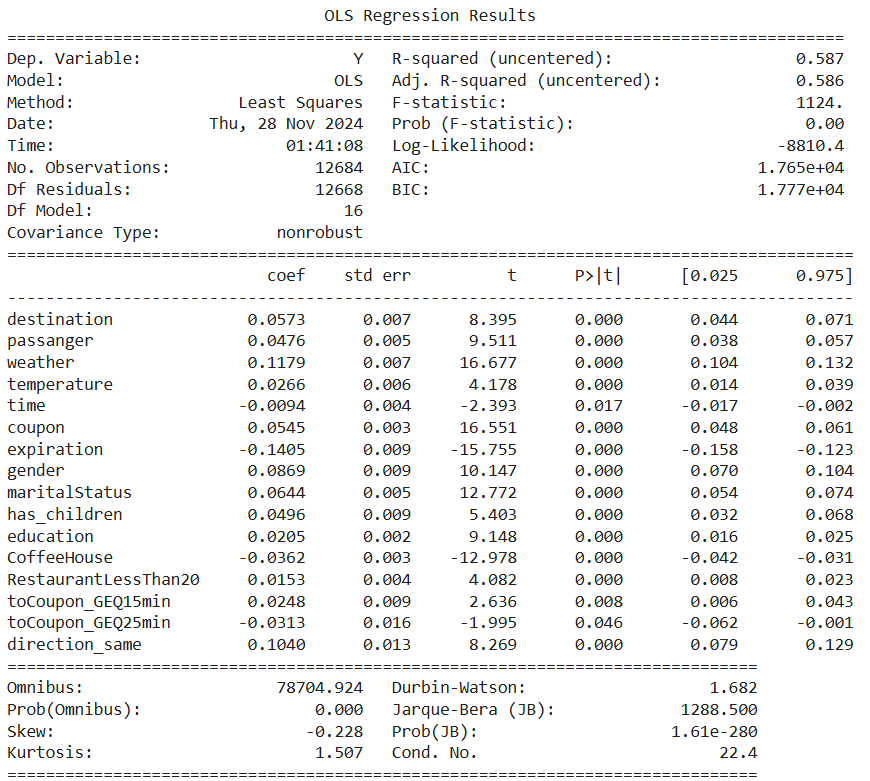


Figure 43: All coefficients are statistically significant

Therefore, age, income, Restaurant20To50, CarryAway and Bar features are dropped to create a dataset that is dimensionally reduced due to stepwise regression. This leads to a dataset with 16 features.

**Forward feature selection**

Python was used to do forward feature selection. 16 features were selected for forward feature selection because stepwise regression resulted with a dataset with 16 features. This was done for consistency. The focus was on reducing the negative mean squared error. Forward feature selection resulted in a dataset where maritalStatus, Bar, RestaurantLessThan20, temperature and direction\_same features were dropped.

Stepwise Regression and Forward Feature selection for dimensionality reduction gave different results resulting in different columns being dropped.

**Backward feature elimination**

Python was used to do backward feature elimination. 16 features were selected for backward feature elimination because stepwise regression resulted with a dataset with 16 features. This was done for consistency. The focus was on reducing the negative mean squared error. Backward feature elimination resulted in a dataset where maritalStatus, Bar, RestaurantLessThan20, temperature and direction\_same features were dropped.

**It has hence observed that forward feature selection and backward feature elimination resulted in the same datasets. Thus, there are 2 dimensionally reduced datasets to work with: the dataset reduced by stepwise regression and the dataset reduced by forward feature selection/backward feature elimination. Dimensionality reduction is the major contribution of towards this dataset compared to past work done on this dataset.**

# Classification Algorithms and Cross Validation

The performance of 5 classification algorithms was evaluated over each of the 2 dimensionally reduced datasets:

1. Random Forest
2. Logistic Regression
3. k Nearest Neighbours (k-NN)
4. Naïve Bayes
5. Decision Tree

**Cross Validation was used and the dataset was split into test and training sets. 20% of dataset was set for testing each time for consistency. The classification algorithms were trained over the training part of the datasets and then evaluated over the testing part of the datasets. Accuracy, classification report and Area under Curve (AUC) were gathered each time.**

Result of Random Forest on stepwise regression reduced dataset:

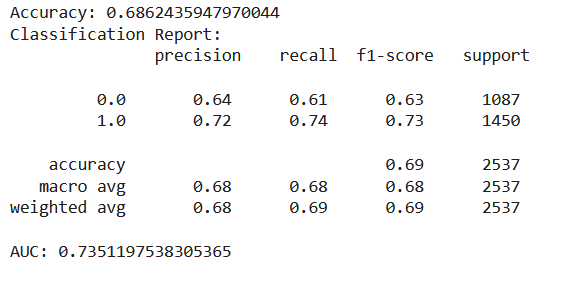


Figure 44: Result of Random Forest on stepwise regression reduced dataset

Result of Random Forest on forward feature selection/backward feature elimination reduced dataset:

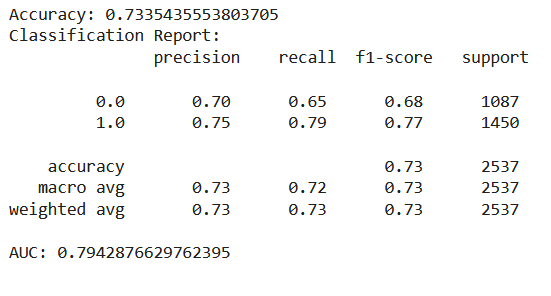


Figure 45: Result of Random Forest on forward feature selection/backward feature elimination reduced dataset

Result of Logistic Regression on stepwise regression reduced dataset:

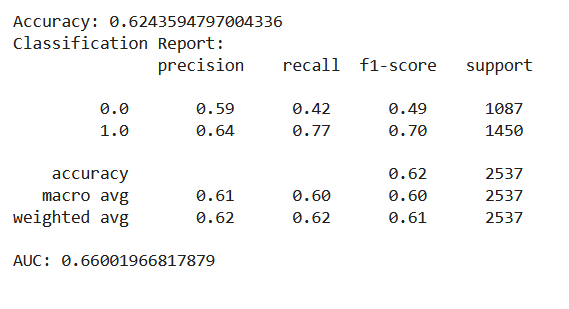


Figure 46: Result of Logistic Regression on stepwise regression reduced dataset

Result of Logistic Regression on forward feature selection/backward feature elimination reduced dataset:

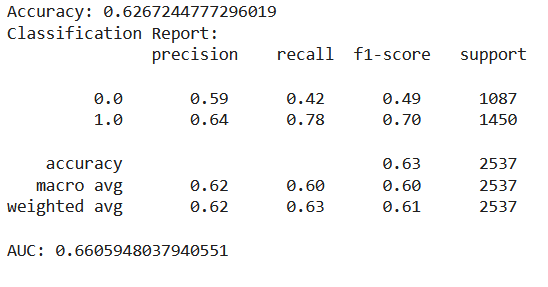


Figure 47: Result of Logistic Regression on forward feature selection/backward feature elimination reduced dataset

Result of k Nearest Neighbours (k-NN) on stepwise regression reduced dataset:

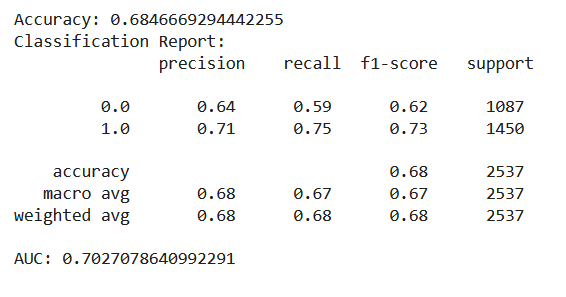


Figure 48: Result of k Nearest Neighbours (k-NN) on stepwise regression reduced dataset

Result of k Nearest Neighbours (k-NN) on forward feature selection/backward feature elimination reduced dataset:

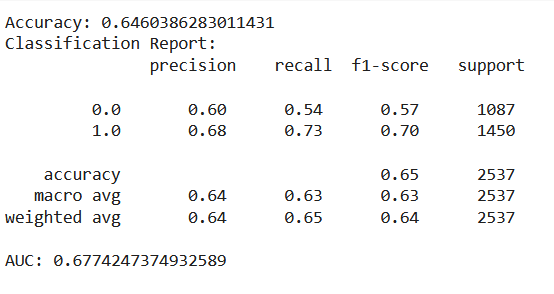


Figure 49: Result of k Nearest Neighbours (k-NN) on forward feature selection/backward feature elimination reduced dataset

Result of Naïve Bayes on stepwise regression reduced dataset:

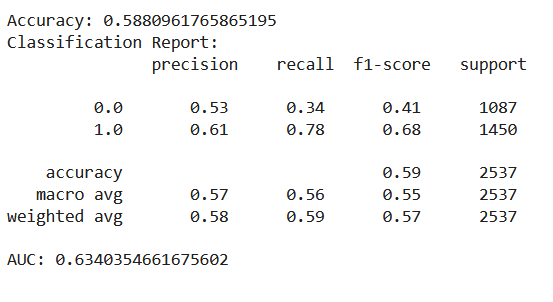


Figure 50: Result of Naïve Bayes on stepwise regression reduced dataset

Result of Naïve Bayes on forward feature selection/backward feature elimination reduced dataset:

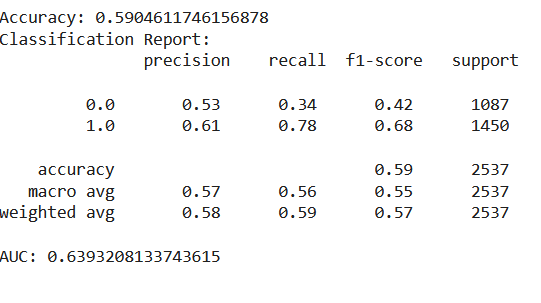


Figure 51: Result of Naïve Bayes on forward feature selection/backward feature elimination reduced dataset

Result of Decision Tree on stepwise regression reduced dataset:

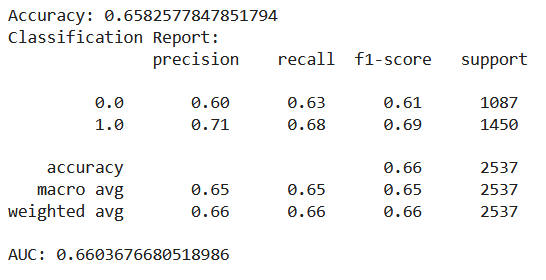


Figure 52: Result of Decision Tree on stepwise regression reduced dataset

Result of Decision Tree on forward feature selection/backward feature elimination reduced dataset:

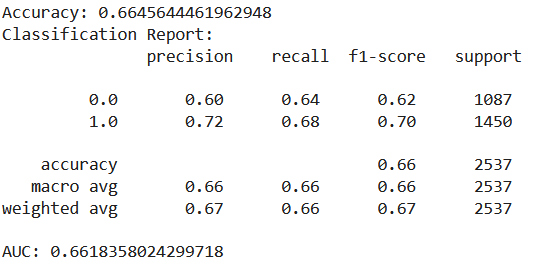


Figure 53: Result of Decision Tree on forward feature selection/backward feature elimination reduced dataset

# Comparison and Analysis of results

Accuracy tells what percentage of the predictions are correct (Ofir Shalev (@ofirdi), 2021). “Precision and Recall are often in tension. That is, improving Precision typically reduces Recall and vice versa” (Ofir Shalev (@ofirdi), 2021). “F1 score combines Recall and Precision to one performance metric. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account” (Ofir Shalev (@ofirdi), 2021). The Receiver Operating Characteristics (ROC) curve is created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings (Ofir Shalev (@ofirdi), 2021). AUC (Area Under Curve):The model’s performance is determined by looking at the area under the ROC curve (or AUC) (Ofir Shalev (@ofirdi), 2021).

Accuracy and AUC was compared for the 5 classification algorithms run over the 2 dimensionally reduced dataset.

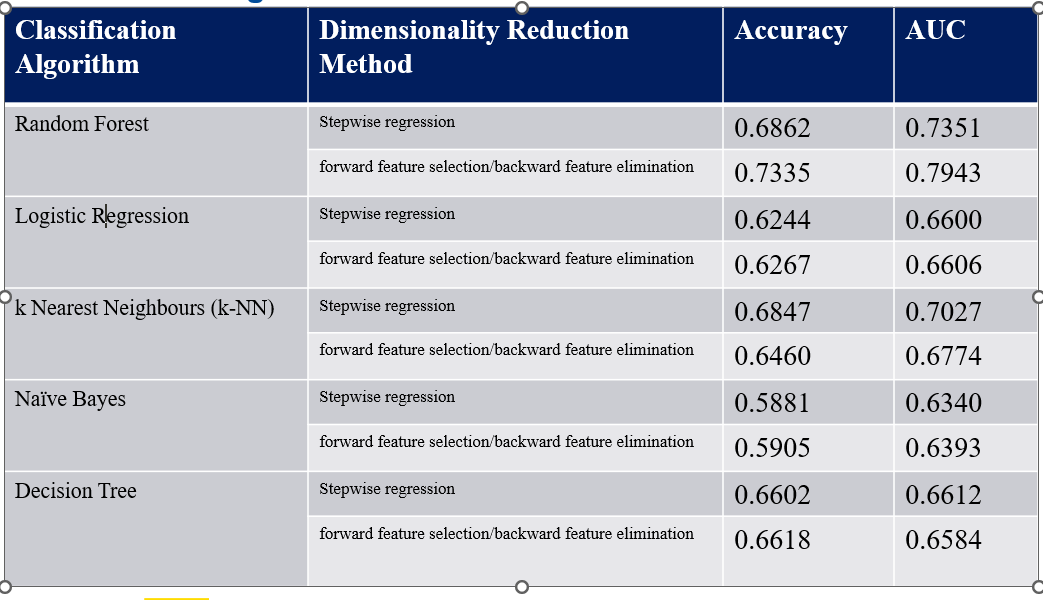


Table 3: Accuracy and AUC on Dimensionally reduced data using different classification algorithms

Random forest on dataset reduced using forward feature selection/backward feature elimination had the highest accuracy and AUC value closest to 1. The Area Under Curve (AUC) close to 1, shows the high predictive power. Whereas Naïve Bayes on Stepwise Regression reduced dataset had lowest accuracy and AUC value closest to 0.5. An AUC value closer to 0.5 shows that it is as good as random chance.

# Methodology

The methodology being followed is based on the video by Babaoglu. (2018, January 6).

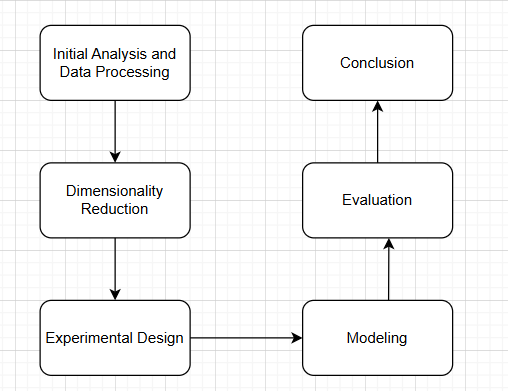


Figure 54: Proposed methodology for this project

In the Initial Analysis, the data will be described and examined. Univariate Analysis will be done to look at the distribution of each attribute. Check for missing values will be conducted. Data Processing will also be done and data will be cleaned.

During dimensionality reduction, a dataset with fewer dimensions using these 3 methods: Stepwise Regression, Forward Feature Selection and Backward Feature Elimination will be acquired.

Experimental Design will involve splitting data into a training set and test set to run classification algorithms over.

Modeling will involve running Random Forest, Decision Tree, Logistic Regression, Naïve Bayes, k-Nearest Neighbours (k-NN) on the dataset.

Evaluation of various classification algorithms will be done by comparing evaluation metrics like the Accuracy, Precision and Area under the Curve (AUC) of each algorithm.

A conclusion is arrived at based on results.

# Limitations / Challenges / Continuity

* Most of the Car feature had missing values and hence the feature could not be used. This is in-vehicle coupon recommendation, hence maybe Car feature was critical in determining whether the coupon would be accepted or not. After all, there are driving scenarios involved.
* This dataset is partially balanced. The results of supervised learning algorithms used to make predictions would skew slightly towards the class with the class with higher percentage of records. The percentage of accepted coupons: 56.843%. The percentage of rejected coupons: 43.157%. In the future, Synthetic Minority Over-sampling Technique (SMOTE) can be used to generate a more balanced dataset.
* The dataset was more focused on a particular type of population. The dataset should have been created by sampling all types of population. For example, Depari et al. (2022) found that the data contained mostly married females who like to travel alone on a sunny day around 6 PM. Most of them have attended college, yet didn’t graduate (Depari et al., 2022). For those who have an occupation, it states that most of them earn an income of around $25000 - $37499 (Depari et al., 2022). It was also mentioned that the destination is mostly the No Urgent Place such as Coffee House, which provides a coupon that expires in one day (Depari et al., 2022).
* Patil et al. (2019) also observed that customers tend to purchase the same coupon over and over again. The dataset for in-vehicle coupon response is deficient in data over a periodic basis to help uncover such patterns. This is a limitation.

# GitHub Repository

<https://github.com/suchetasikdar1/CIND820>

# References

Ahmed, N., & Umair, M. (2024). *Churn prediction using machine learning: A coupon optimization technique*. In World Journal of Advanced Engineering Technology and Sciences (Vols. 12–12, Issue 02, pp. 332–354) [Journal-article]. <https://wjaets.com/sites/default/files/WJAETS-2024-0310.pdf>

Atiq, R., Fariha, F., Mahmud, M., Yeamin, S. S., Rushee, K. I., & Rahim, S. (2022, October 8). *A comparison of missing value imputation techniques on coupon acceptance prediction*. MECS PRess. https://www.mecs-press.org/ijitcs/ijitcs-v14-n5/IJITCS-V14-N5-2.pdf

Babaoglu C. (n.d.). CMTH 642 – Data Analytics: Advanced Methods course [PowerPoint slide printouts]

Babaoglu C. (2018, January 6). *How to conduct data analysis process Systematically* [Video]. YouTube. https://www.youtube.com/watch?v=NMffLbFql5k

Inyama C. (2023). *In-Vehicle-Coupon-Recommendations/coupon\_Project (2).ipynb at main · chrisinyama/In-Vehicle-Coupon-Recommendations*. GitHub. <https://github.com/chrisinyama/In-Vehicle-Coupon-Recommendations/blob/main/coupon_Project%20(2).ipynb>

Depari, G. S., Shu, E., Fachriza, C. A., Chow, J., Wijaya, J., & Winata, R. (2022). Customer’s responses towards in-vehicle coupon recommendation an implementation of business analytics concept. *Jurnal Ekonomi* (Vol. 11, Issue 02) [Journal-article] <https://download.garuda.kemdikbud.go.id/article.php?article=2994923&val=22616&title=CUSTOMER%27S%20RESPONSES%20TOWARDS%20IN-VEHICLE%20COUPON%20RECOMMENDATION%20AN%20IMPLEMENTATION%20OF%20BUSINESS%20ANALYTICS%20CONCEPT>

Niralidedaniya. (2023, January 19). In Vehicle Coupon Recommendation — a Machine learning classification case study. *Medium*. <https://medium.com/@niralidedaniya/in-vehicle-coupon-recommendation-a-machine-learning-classification-case-study-df67e7835703>

Patil, Y., Pawar, O., & Ingle, D. R. (2019). Coupon Purchase Prediction using Machine Learning. In *IJSRD - International Journal for Scientific Research & Development* (Vol. 7, Issue 02) [Journal-article]. <https://ijsrd.com/articles/IJSRDV7I21146.pdf>

*Supervised learning*. (n.d.). Scikit-learn. <https://scikit-learn.org/stable/supervised_learning.html>

UCI Machine Learning Repository. (2020). <https://archive.ics.uci.edu/dataset/603/in+vehicle+coupon+recommendation>

Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017). A Bayesian framework for learning rule sets for interpretable classification. In Maya Gupta (Ed.), *Journal of Machine Learning Research* (Vols. 1–37) [Journal-article]. <https://jmlr.csail.mit.edu/papers/volume18/16-003/16-003.pdf>

*How to run Machine Learning algorithms in GPU*. (n.d.). Stack Overflow. <https://stackoverflow.com/questions/72985935/how-to-run-machine-learning-algorithms-in-gpu>

Ofir Shalev (@ofirdi). (2021, December 10). Recall, Precision, F1, ROC, AUC, and everything - The Startup - Medium. *Medium*. <https://medium.com/swlh/recall-precision-f1-roc-auc-and-everything-542aedf322b9>