CIND 820 Capstone Project

Final Presentation: In-Vehicle Coupon Recommendation

Sucheta Sikdar (ID: 501074722)

Toronto
Metropolitan
University

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).

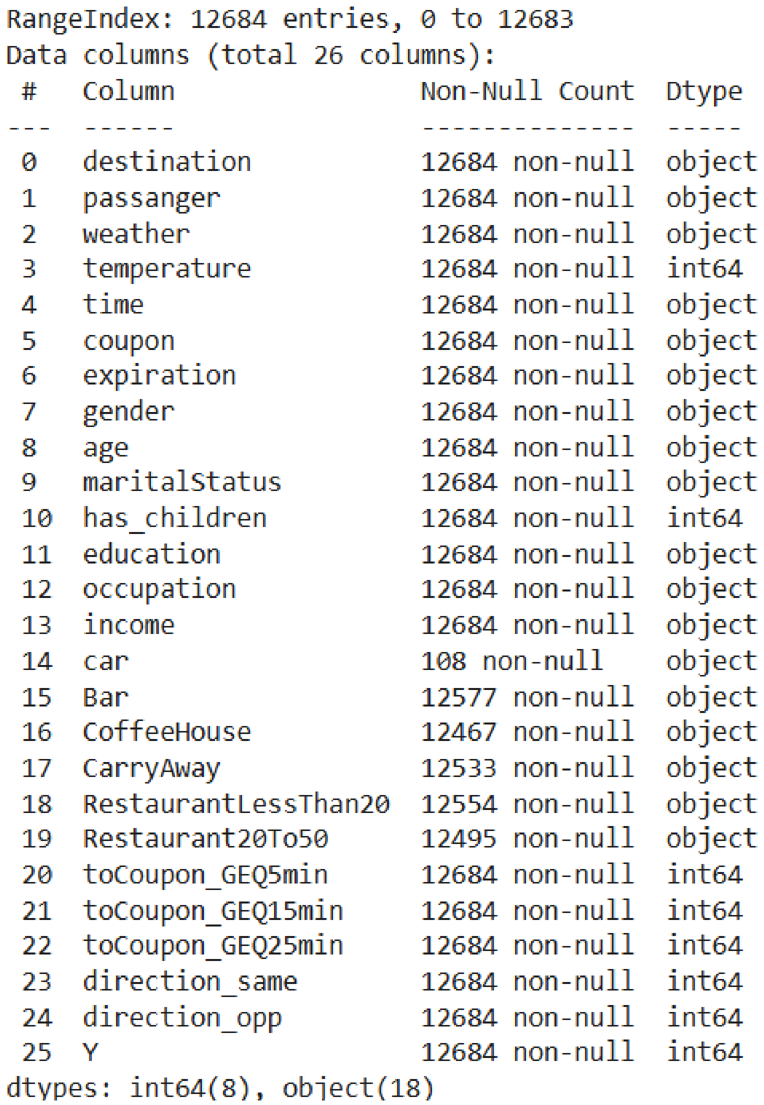
There exists a publicly available dataset called In-Vehicle Coupon Recommendation at UCI Machine Learning Repository from 2020 which describes different driving scenarios of multiple clients and whether the coupon is accepted (UCI Machine Learning Repository, 2020).

Objectives

- 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. 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.

Data Description and Data Preparation

Using Python to yield information on the data:

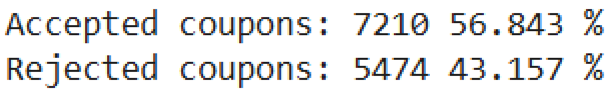


```
RangeIndex: 12684 entries, 0 to 12683
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   destination           12684 non-null  object
 1   passanger             12684 non-null  object
 2   weather               12684 non-null  object
 3   temperature           12684 non-null  int64
 4   time                  12684 non-null  object
 5   coupon                12684 non-null  object
 6   expiration            12684 non-null  object
 7   gender                12684 non-null  object
 8   age                   12684 non-null  object
 9   maritalStatus         12684 non-null  object
 10  has_children          12684 non-null  int64
 11  education             12684 non-null  object
 12  occupation            12684 non-null  object
 13  income                12684 non-null  object
 14  car                   108 non-null    object
 15  Bar                   12577 non-null  object
 16  CoffeeHouse           12467 non-null  object
 17  CarryAway             12533 non-null  object
 18  RestaurantLessThan20  12554 non-null  object
 19  Restaurant20To50      12495 non-null  object
 20  toCoupon_GEQ5min      12684 non-null  int64
 21  toCoupon_GEQ15min     12684 non-null  int64
 22  toCoupon_GEQ25min     12684 non-null  int64
 23  direction_same        12684 non-null  int64
 24  direction_opp         12684 non-null  int64
 25  Y                     12684 non-null  int64
dtypes: int64(8), object(18)
```

Missing values

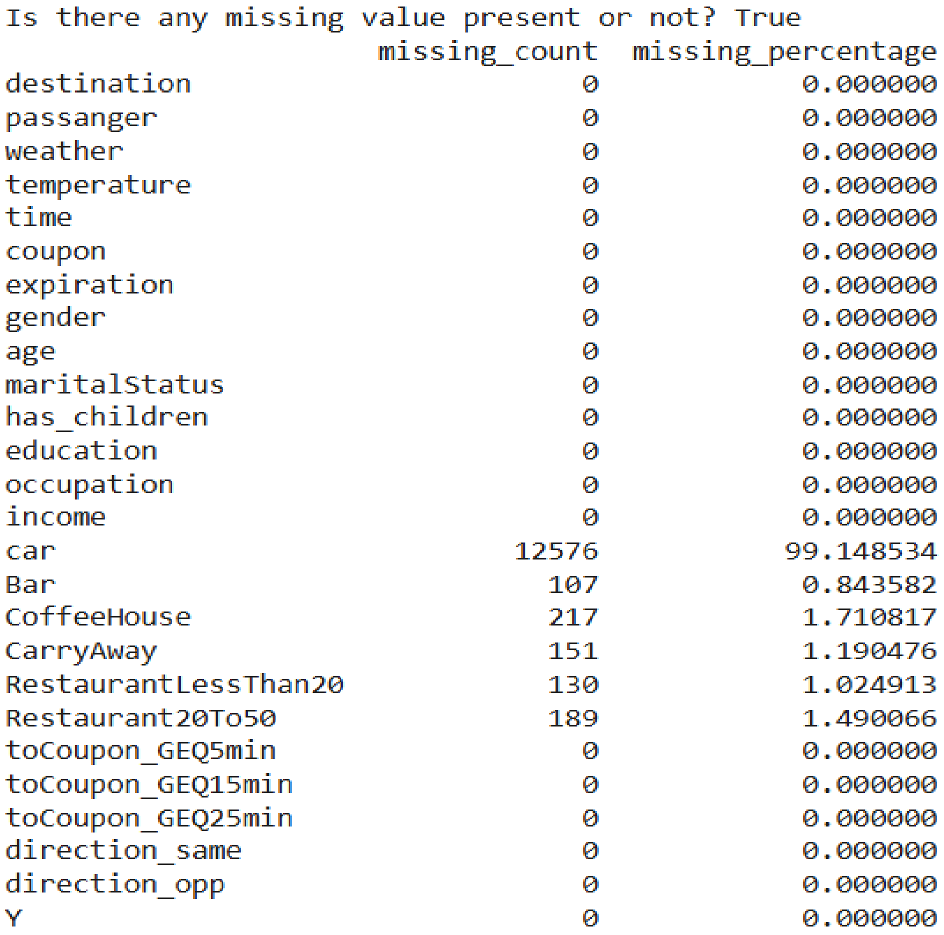
Data Description and Data Preparation contd.

Distribution of target class found using Python:



```
Accepted coupons: 7210 56.843 %
Rejected coupons: 5474 43.157 %
```

Distribution of Missing values in the dataset found using Python:

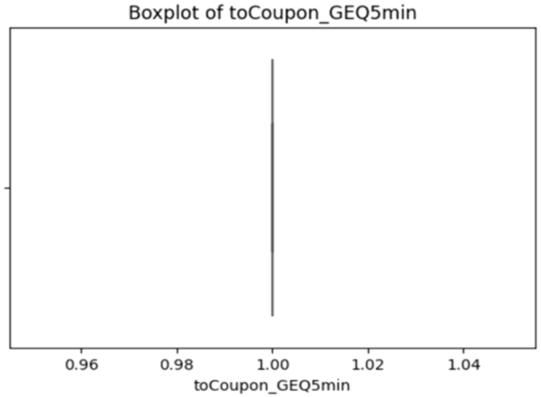


```
Is there any missing value present or not? True
                      missing_count  missing_percentage
destination                       0            0.000000
passanger                         0            0.000000
weather                           0            0.000000
temperature                       0            0.000000
time                              0            0.000000
coupon                            0            0.000000
expiration                        0            0.000000
gender                            0            0.000000
age                               0            0.000000
maritalStatus                     0            0.000000
has_children                      0            0.000000
education                         0            0.000000
occupation                        0            0.000000
income                            0            0.000000
car                           12576           99.148534
Bar                             107            0.843582
CoffeeHouse                     217            1.710817
CarryAway                       151            1.190476
RestaurantLessThan20            130            1.024913
Restaurant20To50                189            1.490066
toCoupon_GEQ5min                  0            0.000000
toCoupon_GEQ15min                 0            0.000000
toCoupon_GEQ25min                 0            0.000000
direction_same                    0            0.000000
direction_opp                     0            0.000000
Y                                 0            0.000000
```

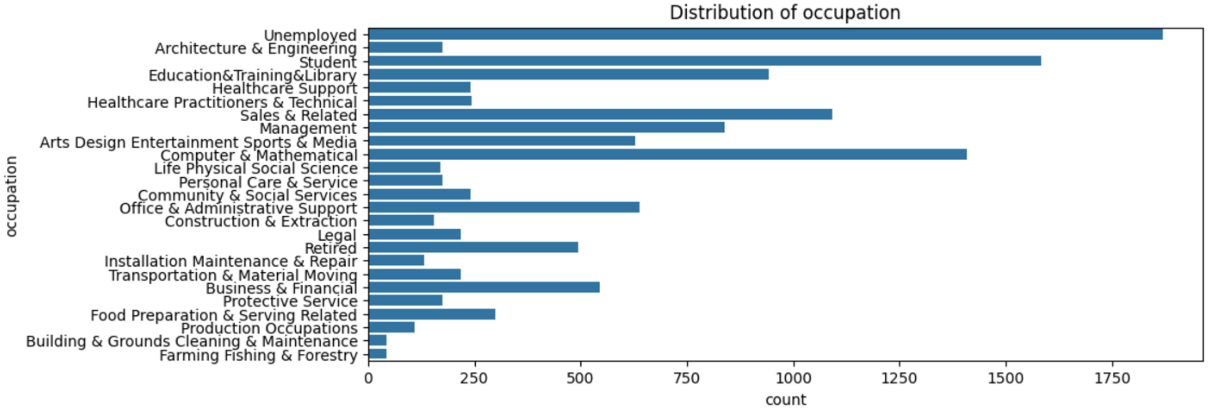
Data Description and Data Preparation contd.

- Dropped column Car as it has too many missing values

- Used Python to generate box plots for numerical variables. Dropped column toCoupon_GEQ5min because there is no variability in its value

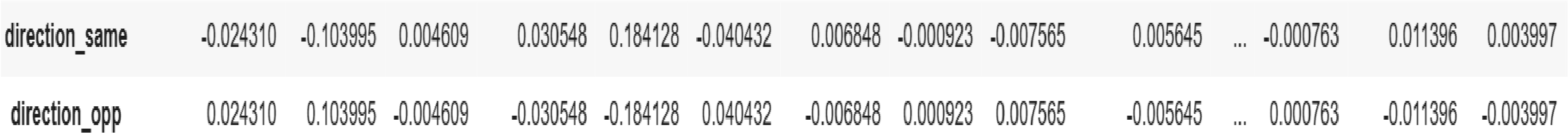


- Used Python to observe the distribution of categorical variables. Dropped column occupation because it has too many categories that leads to a lot of noise.



Data Description and Data Preparation contd.

- Frequent value / mode imputation for missing values in data. 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 used frequent value imputation too for the remaining missing values.
- Did a covariance matrix using Python. Since direction_same has same covariance values as direction_opp. It makes sense to just have one on them and reduce the noise.

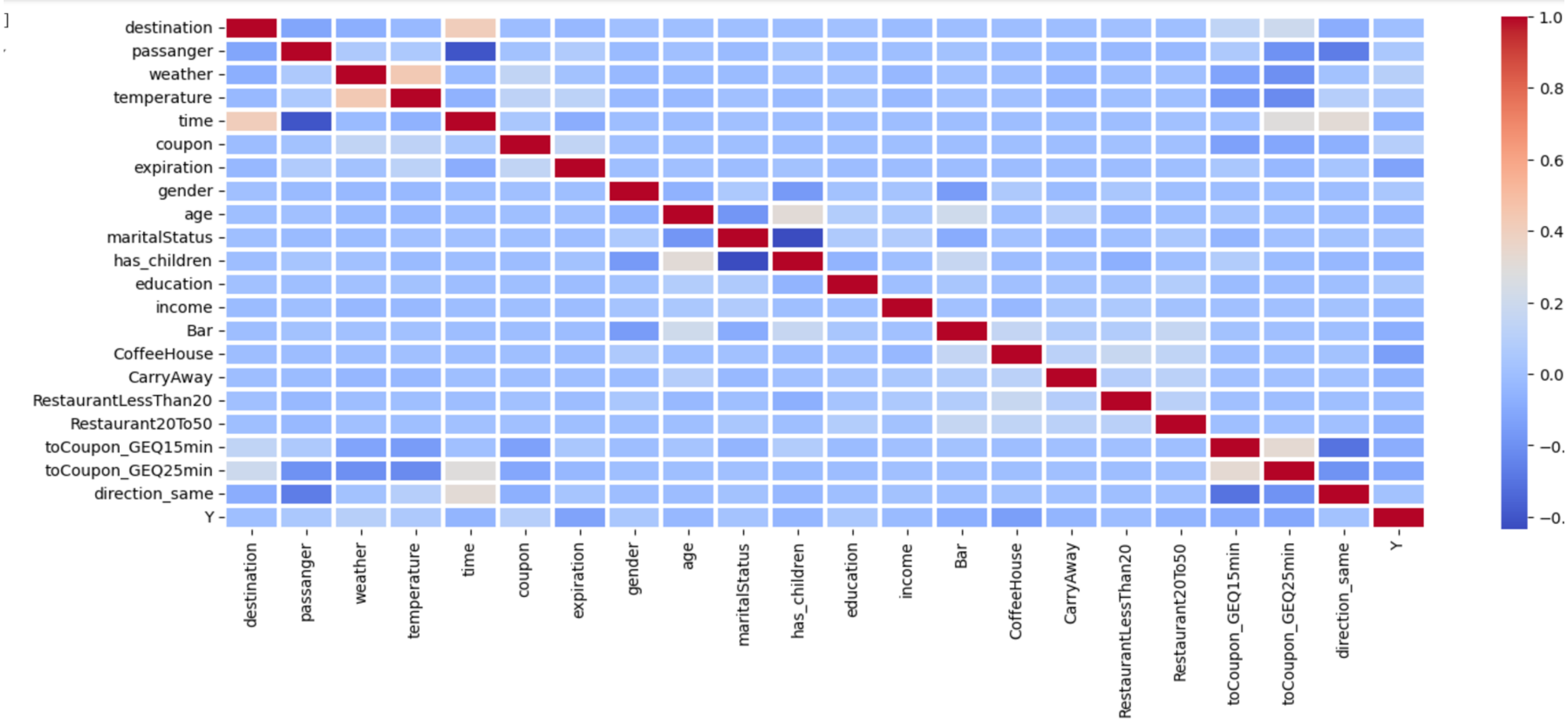


direction_same	-0.024310	-0.103995	0.004609	0.030548	0.184128	-0.040432	0.006848	-0.000923	-0.007565	0.005645	...	-0.000763	0.011396	0.003997
direction_opp	0.024310	0.103995	-0.004609	-0.030548	-0.184128	0.040432	-0.006848	0.000923	0.007565	-0.005645	...	0.000763	-0.011396	-0.003997

Data Description and Data Preparation contd.

- Used a Heat map on correlation matrix to visually catch high correlation.

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.



Dimensionality Reduction

Stepwise Regression

I choose alpha to be 0.05. Coefficients having a p-value of 0.05 or less will be statistically significant.

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

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.

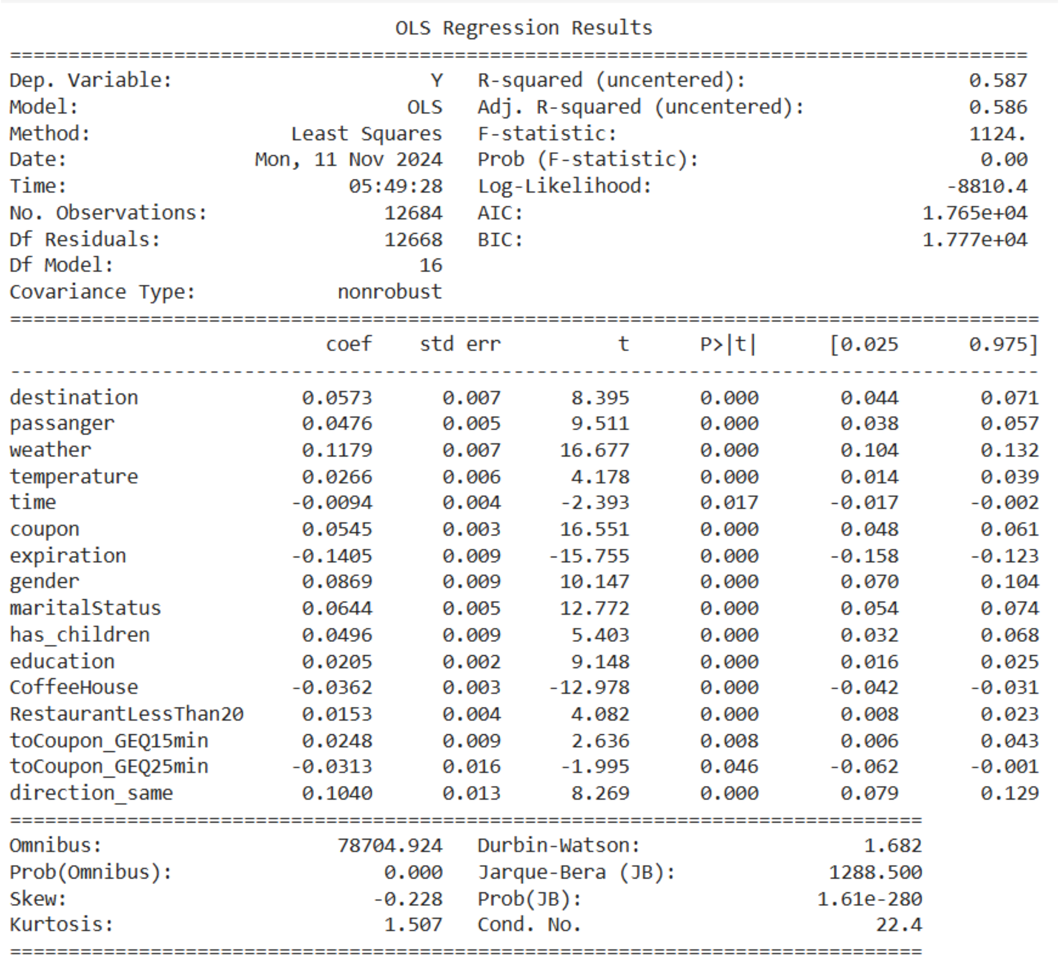
Iteration 3: Restaurant20To50 is least statistically significant because it has highest p-value of 0.576 which is greater than alpha of 0.05. Restaurant20To50 should be dropped.

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.

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.

Dimensionality Reduction contd.

After 5 iterations, 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.



```
                                 OLS Regression Results
=======================================================================================
Dep. Variable:                      Y   R-squared (uncentered):                   0.587
Model:                            OLS   Adj. R-squared (uncentered):              0.586
Method:                 Least Squares   F-statistic:                              1124.
Date:                Mon, 11 Nov 2024   Prob (F-statistic):                        0.00
Time:                        05:49:28   Log-Likelihood:                         -8810.4
No. Observations:               12684   AIC:                                  1.765e+04
Df Residuals:                   12668   BIC:                                  1.777e+04
Df Model:                          16
Covariance Type:            nonrobust
=========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
destination              0.0573      0.007      8.395      0.000       0.044       0.071
passanger                0.0476      0.005      9.511      0.000       0.038       0.057
weather                  0.1179      0.007     16.677      0.000       0.104       0.132
temperature              0.0266      0.006      4.178      0.000       0.014       0.039
time                    -0.0094      0.004     -2.393      0.017      -0.017      -0.002
coupon                   0.0545      0.003     16.551      0.000       0.048       0.061
expiration              -0.1405      0.009    -15.755      0.000      -0.158      -0.123
gender                   0.0869      0.009     10.147      0.000       0.070       0.104
maritalStatus            0.0644      0.005     12.772      0.000       0.054       0.074
has_children             0.0496      0.009      5.403      0.000       0.032       0.068
education                0.0205      0.002      9.148      0.000       0.016       0.025
CoffeeHouse             -0.0362      0.003    -12.978      0.000      -0.042      -0.031
RestaurantLessThan20     0.0153      0.004      4.082      0.000       0.008       0.023
toCoupon_GEQ15min        0.0248      0.009      2.636      0.008       0.006       0.043
toCoupon_GEQ25min       -0.0313      0.016     -1.995      0.046      -0.062      -0.001
direction_same           0.1040      0.013      8.269      0.000       0.079       0.129
==============================================================================
Omnibus:                    78704.924   Durbin-Watson:                   1.682
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1288.500
Skew:                          -0.228   Prob(JB):                    1.61e-280
Kurtosis:                       1.507   Cond. No.                         22.4
==============================================================================
```

Dimensionality Reduction contd.

- Forward Feature Selection
- Backward Feature Elimination

I wanted to keep the same number of features as the dataset reduced by stepwise regression, so I picked 16 features for generating dataset generated by forward feature selection and dataset generated by backward feature elimination. I wanted consistency.

It was observed that the results of Forward Feature Selection and Backward Feature Elimination are the same and give the same reduced dataset.

Thus, there are 2 dimensionally reduced datasets:. the one reduced using stepwise regression and the other reduced using backward feature elimination/forward feature selection.

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.

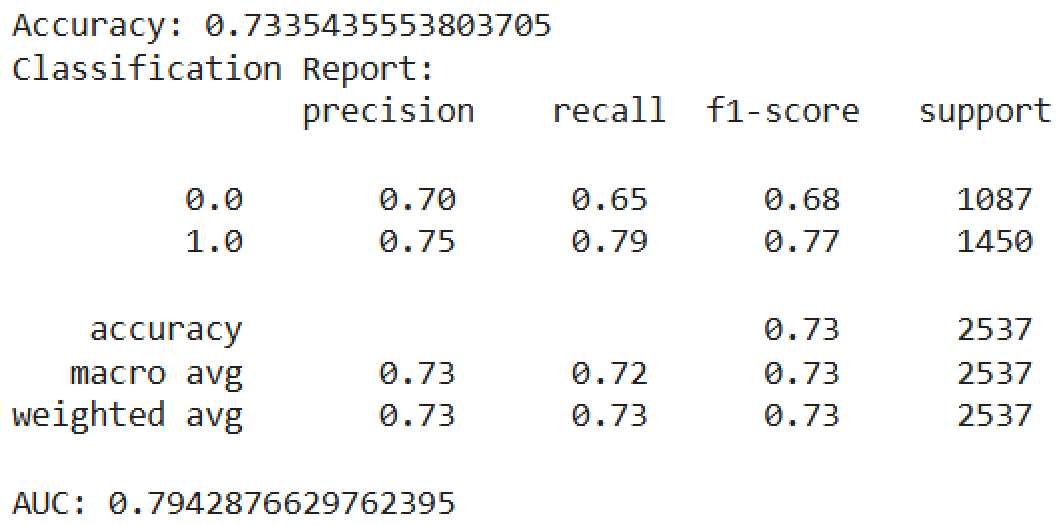
Accuracy and AUC on Dimensionally reduced data using different classification algorithms

Classification Algorithm	Dimensionality Reduction Method	Accuracy	AUC
Random Forest	Stepwise regression	0.6862	0.7351
	forward feature selection/backward feature elimination	0.7335	0.7943
Logistic Regression	Stepwise regression	0.6244	0.6600
	forward feature selection/backward feature elimination	0.6267	0.6606
k Nearest Neighbours (k-NN)	Stepwise regression	0.6847	0.7027
	forward feature selection/backward feature elimination	0.6460	0.6774
Naïve Bayes	Stepwise regression	0.5881	0.6340
	forward feature selection/backward feature elimination	0.5905	0.6393
Decision Tree	Stepwise regression	0.6602	0.6612
	forward feature selection/backward feature elimination	0.6618	0.6584

Interpretability and Insights

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.

Generated by Python for Random Forest on the dataset reduced by forward feature selection/backward feature elimination:



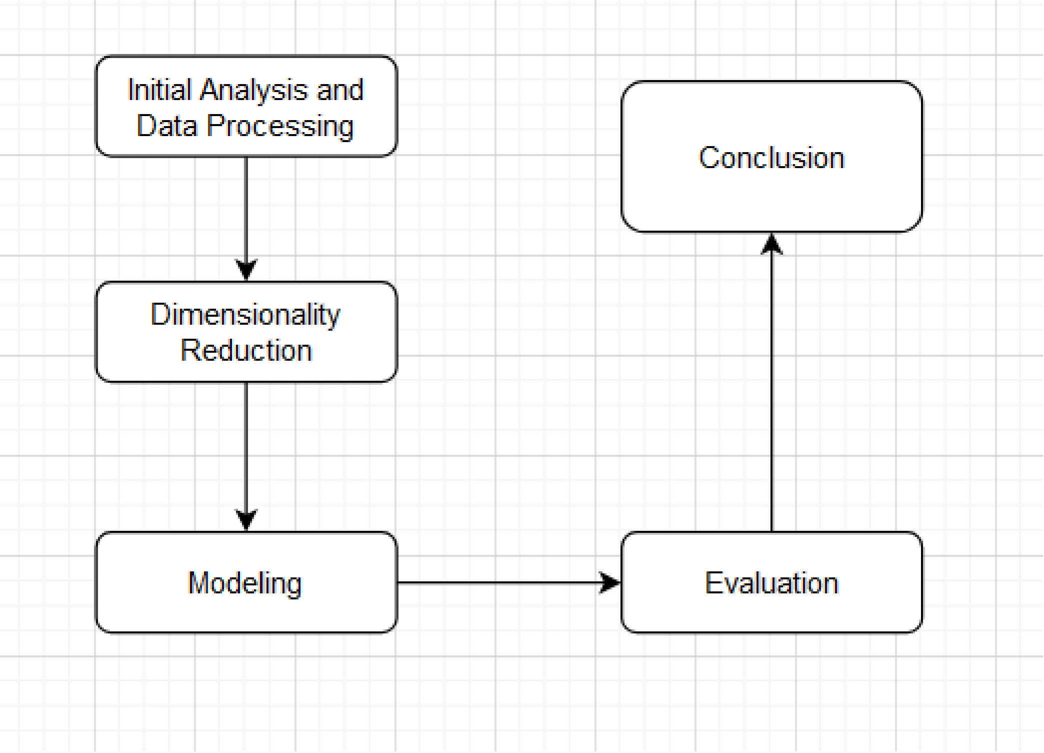
```
Accuracy: 0.7335435553803705
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.65      0.68      1087
         1.0       0.75      0.79      0.77      1450

    accuracy                           0.73      2537
   macro avg       0.73      0.72      0.73      2537
weighted avg       0.73      0.73      0.73      2537

AUC: 0.7942876629762395
```

Methodology



Initial Analysis and Data Processing

Dimensionality Reduction

Modeling

Evaluation

Conclusion

Toronto Metropolitan University

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.
- 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 Page 13 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.

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=CUST OMER%27S%20RESPONSES%20TOWARDS%20IN Page 43 VEHICLE%20COUPON%20RECOMMENDATION%20AN%20IMPLEMENTATION%20OF %20BUSINESS%20ANALYTICS%20CONCEPT

Toronto Metropolitan University

References contd.

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

Thank you. Questions are welcome

Toronto Metropolitan University