Machine Learning: Predicting in Vehicle Coupon Acceptance Rates Using Different Models

Wong Jing Hei (UID: 805643634)

***Abstract*— This paper uses data surveyed through Amazon Mechanical Turk on in-vehicle coupon recommendations that is sourced from the UCI Machine Learning Repository. A total of 12684 samples were recorded in the dataset, indicating whether an individual had accepted a coupon for a restaurant that expires within 2 hours or a day. With marketing become more selective and targeted, building a model to classify whether an individual is receptive to certain forms of marketing can prove to be very informative. Four different types of models are considered, namely Logistic Regression, Lasso Regression, Ridge Regression and Support Vector Machine (SVM), and are all subsequently evaluated respectively based on their predictive accuracy, model score and confusion matrices. After evaluation, this paper concludes that an SVM with a polynomial kernel offers the best predictive model to classify whether an agent accepts a restaurant coupon or not.**

***Index Terms*—Machine Learning, Multivariate Classification, Logistic Regression, SVM**

I. INTRODUCTION

With advertising becoming more personalized and targeted, data offers new ways for firms to analyze customer behavior and adjust their marketing strategy accordingly in order to maximize efficiency and profits. One of the ways firms can do this is by generating a model that can predict whether an individual is likely to accept or reject a marketing offer based on his demographic or other features. To test and compare whether such a model is feasible, I use advertisement data surveyed through Amazon Mechanical Turk on in-vehicle coupon recommendations acceptance rates to try and compare the predictive accuracy across 4 different types of classification models.

II. TASK DESCRIPTION

Using advertisement data surveyed by Amazon Mechanical Turk, 4 different classification models will be fit over the specified data.

A. Model 1: Logistic Regression

The first classification model is a **Logistic Regression**, and is the one of the simplest forms of classification models used to make predictions, with a binary Y variable that is labelled between 0 and 1. In this case, the Y variable will be equal to 1 if it the customer accepts the coupon, and 0 if the customer does not. A logistic model will be trained using training data before being fitted over testing data to test its out of sample predictive accuracy, as well as other evaluation metrics.

B. Model 2: Lasso Regression

The second classification model is a **Lasso Regression**, which works similarly with a Logistic Regression, but adds a penalty term to the loss function to prevent overfitting of the training data. This essentially means a penalty term is introduced so that the specified model does not become too good at predicting data from the training set, but poor at predicting other data. By allowing the data to be slightly worse fit, such a model would be better at making long term predictions. The penalty term for a lasso regression is equal to lambda times the absolute value of the slope, and is added to the loss function as a joint function to be minimized.

C. Model 3: Ridge Regression

The third classification model is a **Ridge Regression**, which works almost identically as a Lasso Regression, but instead adds a penalty term to

the loss function that is equal to lambda times the squared magnitude of the slope. The joint loss function of the model is then minimized to prevent overfitting of the model.

D. Model 4: Support Vector Machine

The fourth and last classification model is through the use of a **Support Vector Machine (SVM)**. An SVM acts as a non-linear classification model used to optimize data with many sample features. SVMs also have different kernel types, as well as high dimension feature spaces that can be chosen accordingly to what fits the data best through cross validation. This allows SVMs to have some degree of flexibility that other models do not, allowing it to operate efficiently.

III. Major Challenges and Solutions

There are additional parameters that need to be chosen in order for models 2,3 and 4 to work. For the Lasso and Ridge Regression, an "alpha" parameter needs to be chosen to multiply the penalty term with to prevent overfitting. For SVM, a c value and kernel type need to be chosen to best fit the data. As the values of these parameters heavily affect the predictive performance of these models, it is important that ideal parameters are chosen. While initially challenging, this can be resolved by comparing the out of sample test performance of the models from a set of initial parameters, then choosing the parameters that lead to models with the best performance. This can ensure that the optimal parameters are chosen.

IV. Experiments

A. Data Description

Given the large number of features recorded in the dataset, a lot of data cleaning have to be done. Many variables in the dataset are categorical variables, including variables such as age, income, occupation, coupon type, expiration date and more. Hence dummy variables need to be created for each recorded category, omitting one from each category to avoid multicollinearity. The large number of features also means that I need to ensure that none of the features are skewed to avoid bias. The following explanatory variables will be used in this study:

1) Destination (Categorical): Intended location of the driver when advertised.
2) Passenger (Categorical): Who the driver is travelling with when coupon is advertised.
3) Weather (Categorical): Weather of day
4) Temperature: Temperature of Day
5) Time: Time Coupon was marketed
6) Coupon Type (Categorical): Type of coupon marketed: e.g., Bar, Restaurant, Coffee House
7) Expiration (Dummy): Equals 1 if coupon expires in a day, equals 0 if coupon expires in 2 hours.
8) Gender (Dummy): Equals 1 if individual is male, equals 0 if female.
9) Age (Categorical): Age of individual
10) Marital Status (Categorical): Unmarried, Single, Married, Divorced
11) Has Children (Dummy): Equals 1 if individual is a parent, equals 0 otherwise.
12) Education (Categorical): Level of Education
13) Occupation (Categorical): occupation of individual
14) Income (Categorical): Level of Income
15) Bar (Categorical): How frequent the individual has visited a bar in the past month.
16) Coffee House (Categorical): How frequent the individual has visited a coffee house in the past month.
17) Carry Away (Categorical): How frequent the individual has visited a take-out place in the past month.
18) Restaurant <$20 (Categorical): How frequent the individual has visited a restaurant with average spending <$20 in the past month.
19) Restaurant $20 to $50 (Categorical): How frequent the individual has visited a restaurant with average spending between $20 to $50 in the past month.
20) toCoupon_GEQ15min (Dummy): Equals 1 if the driving distance to the advertised restaurant is greater than 15 minutes.
21) Direction (Dummy): Equals 1 if the advertised restaurant is in the same direction as the individual's destination.

A few graphs are shown below to illustrate the summary statistics for a few of the key X variables to ensure that the data isn't skewed with a sufficient sample size for each category:

Figure 1: Age Distribution

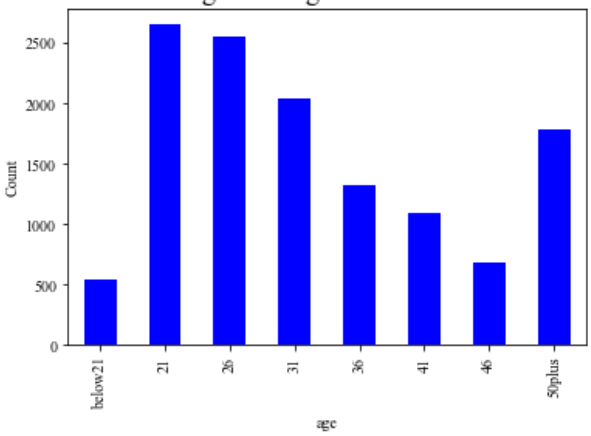


Figure 2: Income Distribution

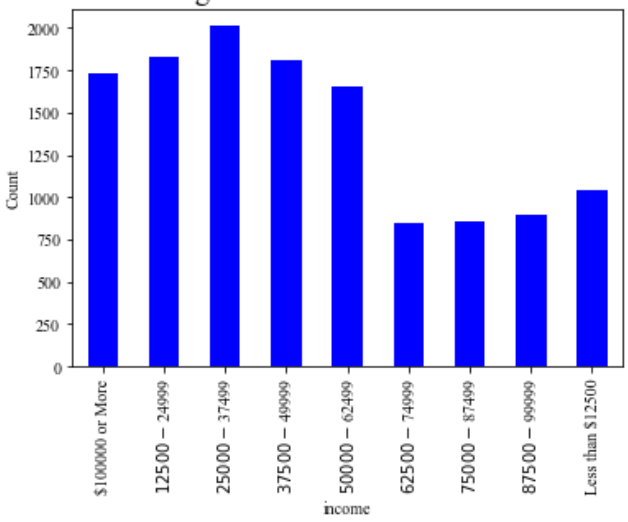
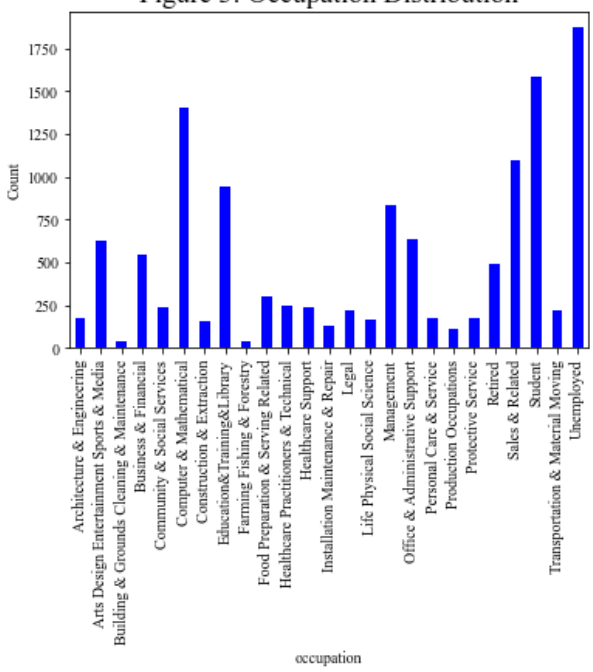


Figure 3: Occupation Distribution



Looking at figure 3, we can observe that there are a total of 25 different recorded types of occupations in the dataset, which seems a bit much to all control for. Given that a lot of this data subdivides employed occupations into further categories, I have instead decided to group occupation into 4 different groups: employed, unemployed, retired and student. This is done to ensure that there is a sufficient sample size for each category. As the distribution seen in figure 1 and figure 2 is relatively more evenly skewed, no extra transformation is required to the data.

Figure 4: Coupon Type

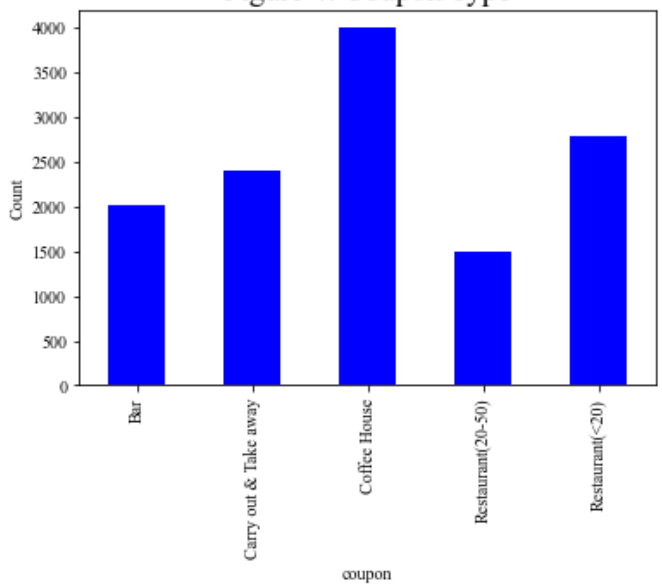
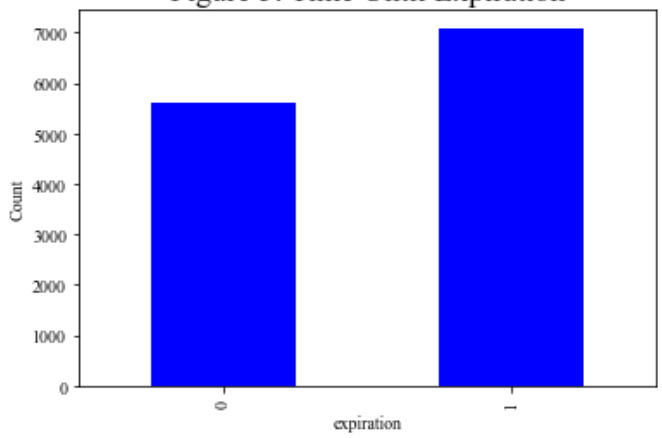


Figure 5: Time Until Expiration

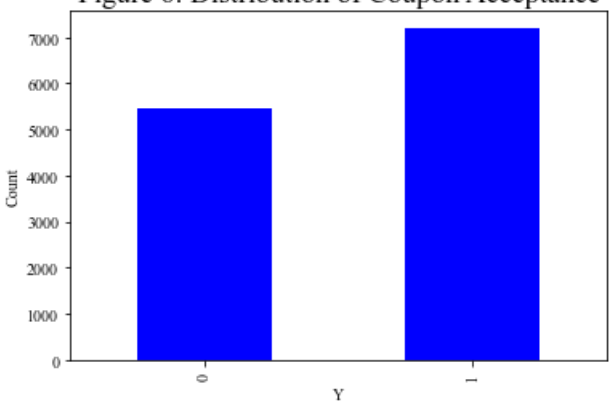


The dataset involves coupons for different types of restaurants. Figure 4 summarizes the distribution of these coupon types, involving bars, takeout places, coffee houses, restaurants with expected spend below $20 and restaurants with expected spend between $20-$50. The coupons in this dataset also only expire in two distinct periods of time, and its distribution can be observed in figure 5, where expiration = 1 if the coupon expires in 1 day and

expiration = 0 if the coupon expires in 2 hours. While these samples are not perfectly evenly distributed, they are sufficiently evenly distributed enough such that there is no strong cause for concern of bias inherent within the data.

Finally, we look at the distribution of the Y variable.

Figure 6: Distribution of Coupon Acceptance



In figure 6, we can see that the distribution of coupon acceptance is pretty even, where Y=1 if the individual accepted the marketed coupon and Y=0 if the individual rejected the marketed coupon, hence sample is sufficiently evenly distributed for unbiased analysis.

To use and test our model, the data is randomly divided into two equal sized sets of training and testing data (6342 each). To ensure that each model is evaluated fairly and that the machine learning process is efficient and consistent, all the data will undergo Min Max Normalization, since a majority of the data involves dummy variables.

Due to the large number of explanatory variables (67 total), there is insufficient room in a page for a correlation plot to be shown to illustrate the correlation between each variable. However, to ensure that multicollinearity is not a strong cause for concern in the sample, a Variance Inflation Factor test was run on all the explanatory variables, and a mean VIF value of 4.60 is found, which is < 10. Hence multicollinearity is not a strong cause for concern.

B. Evaluation Metrics

In the subsequent sections of this paper, I will fit the data over training data using each of the four different models and evaluate its out of sample accuracy, R^2 score, and confusion matrices on the testing data. Confusion matrices are tabulated using the following method:

TABLE 1
CONFUSION MATRIX

Classification	**Actual Positive**	**Actual Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Accuracy is therefore measured by the percentage of correct predictions a specified model makes, calculated by dividing the sum of TP and TN with the total number of observations. Other evaluation metrics can be derived from the confusion matrix that will be used to evaluate the performance of each model. Precision measures the ratio of accurate positive predictions made out of all positive predictions, and is measured by dividing TP by the sum of TP and FP. Recall rate on the other hand measures the ratio of accurate positive predictions out of all positive samples, and is calculated by dividing TP by the sum of TP and FN. Receiver Operating Characteristics (ROC) graphs will also be drawn to visualize each model's aggregate performance. The area under curve (AUC) of the ROC will be also subsequently evaluated to determine model performance. The closer AUC is to 1, the better the model, while the closer AUC is to 0.5, the worse.

C. Major Results

Based on testing the predictive power of Lasso, Ridge Regressions and SVMs over different hyperparameters, results indicate that the Lasso Regression is optimized when an alpha value equal to 10 is specified, while a Ridge Regression is optimized when an alpha value of 0.001 and a "balanced" class weight is specified. SVMs on the other hand is optimized with a c-value equal to 8 and a polynomial kernel type. Since a polynomial kernel is used when SVM is optimized, this suggests that the SVM model will have significantly different levels of performance than logistic regressions as higher hyper dimensions are invoked in the machine learning process, though this does not necessarily mean that the SVM model is definitively superior without further analysis.

To determine which model is the superior model, I evaluate each model using the evaluation metrics specified in part B. To start, table 2 illustrates the model score (R^2) of each of the models across both training and testing data.

TABLE 2
MODEL PERFORMANCE

Model	Model Score	
	Training	Testing
Logistic	0.6963	0.6810
Lasso	0.6966	0.6813
Ridge	0.6935	0.6827
SVM	0.9674	0.7235

Based on the results seen in table 2, it is clear the SVM model is head and shoulders above the other 3 models in terms of predictive power, with a model score of 0.9674 and 0.7235 on the training and testing datasets respectively. This indicates how SVM is the best predictive model to be used when making predictions with the advertisement dataset, as the second closest model score for both sets is merely 0.6966 and 0.6827 respectively.

Table 3 offers further evidence of the strength in model 4 by illustrating the confusion matrix statistics of each model.

TABLE 3
EVALUATION METRICS

Model	Accuracy	Precision	Recall Rate	ROC AUC
Logistic	0.6810	0.6989	0.7683	0.733
Lasso	0.6813	0.6997	0.7669	0.733
Ridge	0.6827	0.7177	0.7257	0.733
SVM	0.7235	0.7507	0.7671	0.785

As illustrated, the SVM model has the highest performance in terms of accuracy, precision and ROC AUC out of all the models. Interestingly, the SVM model performs similarly with other models in terms of recall rate, yet the other metrics still indicate that the SVM model has the highest predictive power out of all the models. It is also important to note that the ROC curves for models 1 to 3 have very similar AUC values. To visualize this phenomenon, figures 7 to 10 will be plotted to compare the predictive power of each model:

Figure 7: ROC Curve for Model 1

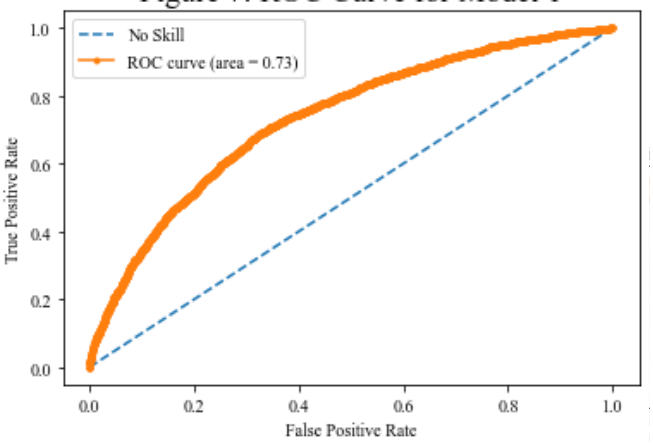


Figure 8: ROC Curve for Model 2

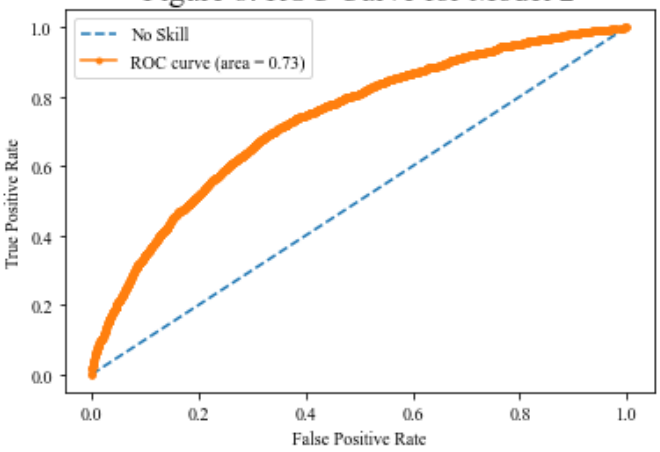


Figure 9: ROC Curve for Model 3

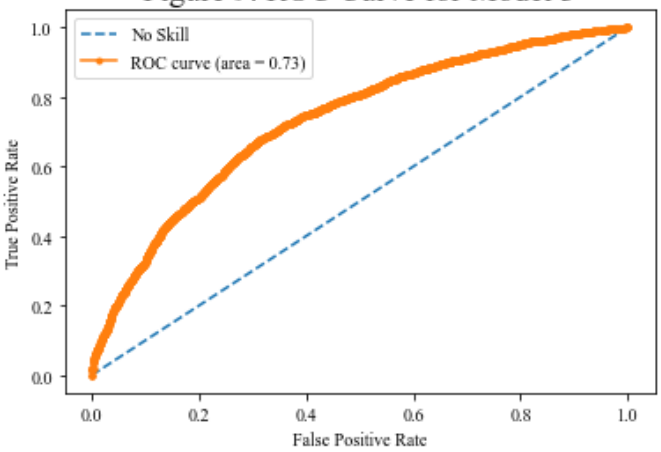
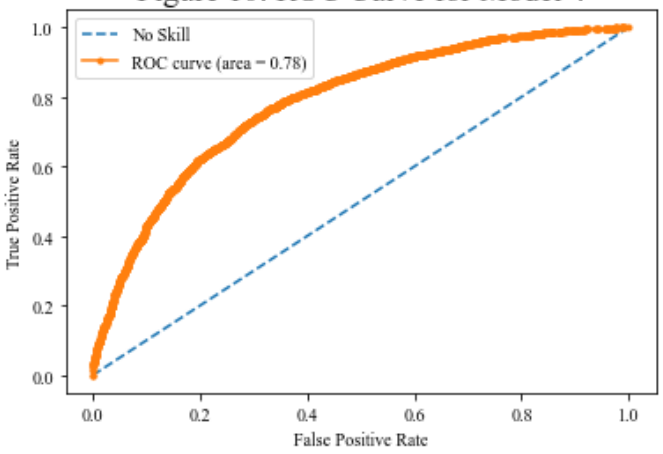


Figure 10: ROC Curve for Model 4



D. Analysis

It is interesting to observe how the model score and AUC values of the Logistic, Lasso and Ridge regressions are very similar across both training and testing data. This could potentially be explained as all three of these models are fundamentally similar in nature based upon logistic regressions, with Lasso and Ridge regressions only adding different penalty terms to each regression to minimize the problem of overfitting. Given how overfitting does not seem to be a predominant problem in the base logistic regression (model 1) as model performance is similar across both training and testing data (0.6963 and 0.6810 respectively), it would make sense that models 1 to 3 are similar in other metrics of predictive power as well. SVMs on the other hand take a significantly different mathematical approach to logistic regressions by maximizing the margins with the closest support vectors to generate predictions. As mentioned before, given how the SVM model is optimized under a polynomial kernel type under this dataset, it is not surprising to see the SVM model to perform significantly differently than the other logistic regression-based models, hence accounting for the significant differences in model scores between SVM and the other models.

V. Conclusion and Future Works

Through the use of different machine learning techniques, this paper fitted and compared 4 different regression models over an advertising dataset sourced from UCI's Machine Learning Repository, which describes in-vehicle restaurant coupon acceptance rates across a random population. After splitting the data into two equal sets of training and testing data, my analysis concluded that an SVM model with a polynomial kernel type best fits the training data and is the most accurate in making out of sample predictions, with high precision rates and ROC AUC values.

However, while the SVM model is decent at making predictions, it is by no means perfect, as there is still a sizable margin of error (>20%) when making predictions based on the model. Possible extensions to this study include developing an artificial neural network to fit the data using a sigmoid or tanh function in order for more accurate classifications to be made.

References

[1] Dalwinder S., Birmohan S. (2020), "Investigating the impact of data normalization on classification performance', Applied Soft Computing, Volume 97, Part B ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2019.105524.

[2] K. Bache and M. Lichman. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml

[3] Salazar, Diego & Velez, Jorge & Salazar Uribe, Juan. (2012). "Comparison between SVM and Logistic Regression: Which One is Better to Discriminate?". Revista Colombiana de Estadística. 35. 223-237.

[4] Wang, T., Rudin, C. et al. (2017), "A Bayesian Framework for Learning Rule Sets for Interpretable Classification". The Journal of Machine Learning Research 18, no. 1. 2357-2393