THE COMPARISONS OF MACHINE LEARNING ALGORITHMS FOR PREDICTING THE ACCURACY OF IN-VEHICLE COUPON RECOMMENDATION

Mehar Jyothi Karpurapu, Janani Suresh Kumar, Nikhilesh

***Abstract* – The research aimed at the case of in-vehicle coupon recommendation for the customers and compares the predictive accuracy among three machine learning algorithms. We apply our method (Logistic Regression, Decision Tree, Random Forest) to characterize and predict behavior of the customer with respect to the in-vehicle recommendation system. Therefore, among the three techniques the Random Forest is the one that is give more accuracy of the data.**

Keywords – Supervised Learning, Random Forest, Logistic Regression, Decision Tree.

INTRODUCTION

The goal of the research paper is to predict whether a particular customer will accept the coupon or not. The dataset used in the project consists of labelled data hence it is a supervised learning method. In supervised learning, we are using classification algorithm for training the model. The dataset has various features which describes the characteristics of the customer. By understanding and analysing the customer context, the response will be predicted.

Firstly, the dataset is cleaned by clearing out the null values. Then the dataset is explored for any similarity or pattern within the columns and combine them accordingly. The dataset is now split into training and testing datasets. For training the dataset, three classification algorithm is used. The accuracy score is calculated for each model.

We have chosen three classification algorithms for training the model.

Logistic Regression: It is a classification machine learning algorithm to predict a data based on prior observations of the dataset. It is a statistical method which analyses the relationship between one or more independent variable and predict the output (dependent variable).

Decision Trees: It is a classification algorithm which predicts the output by learning simple decision rules from data features. Decision trees can be used for both regression and classification problems.

Random Forest: It is a classification algorithm similar to decision trees. In Random Forest Algorithm, the output data is predicted by comparing the outputs by taking mean or average from various decision trees. Random forest can be used for both classification and regression problems.

LITERATURE SURVEY

Wang, Tong, Cynthia Rudin, Finale Doshi-Velez, Yimin Liu, Erica Klampfl, and Perry MacNeille, explained about building classifiers in their paper using short set of rules. Bayesian Rule Sets – BRS method is a machine learning algorithm used to characterize and predict user behaviour according to in vehicle personalized recommendation system. The advantage is that the decision trees in Bayesian method does not greedily grow the model. [1]

M. K. Dahouda and I. Joe, published a research paper explaining about pre-processing categorical data using deep learning. They proposed a new technique called deep learning embedding to encode categorical variables that can be both ordinal or nominal. Each categorical feature is mapped with distinct vector by learning the properties of each vector using neural networks. The accuracy score for deep-learning embedding technique is 89% which is higher than one-hot encoding (71%). [2]

Alexander Smirnov, Nikolay Shilov, published a research paper which explains about the integration of various information services with the help of recommendation system by using the previous decisions made by the driver to collect the similar interest for choosing the appropriate coupon. The algorithm is used when a driver has a lot of coupons, this model helps the driver to choose the appropriate coupon for a particular situation. Driver receives notification when there are any special offers available at nearby stores. [3]

Saeed Mirshekari, wrote a blog on LinkedIn about his first Kaggle competition titled "Coupon purchase prediction". The competition is to find the 10 most likely to get purchased coupons from hundreds of coupons available. He used two approaches: cosine similarity and decision trees. [4]

Cheng Yeh a, Che-hui Lien, compares the predictive accuracy of default credit card payment in Taiwan using six different data mining techniques. There are several data mining techniques used such as Logistic Regression, Naïve Bayesian Classifier, Artificial Neural Networks (ANN) etc. [5]

CLASSIFICATION ACCURACY AMONG ML ALGORITHMS

1. Description of data

Our study took in-vehicle coupon recommendation multivariate data set. It is collected via a survey on Amazon Mechanical Turk and describes different driving scenarios including occupation, income, passenger, destination etc. As a result, it questions the customer that whether he will accept the coupon if he is the driver. It has used the following variables as explanatory variables:

- X1: Destination: It describes the destination of the users in vehicle (No urgent Place, Home, Work)
- X2: Passenger: It describes the passenger details in the vehicle (Alone, Friend(s), Kid(s), Partner)
- X3: Weather (Sunny, Rainy, Snowy)
- X4: Temperature (55, 80, 30)
- X5: Time (2PM, 10AM, 6PM, 7AM, 10PM)
- X6: Coupon (Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50))
- X7: Expiration (1d, 2h (the coupon expires in 1 day or in 2 hours))
- X8: Gender (Female, Male)
- X9: Age (21, 46, 26, 31, 41, 50plus, 36, below21)
- X10: maritialStatus (Unmarried partner, Single, Married partner, Divorced, Widowed)
- X11: has_Children: The value 1 is considered as True and 0 as False. (1, 0)
- X12: education (Some college - no degree, Bachelors degree, Associates degree, High School Graduate, Graduate degree (Masters or Doctorate), Some High School)
- X13: 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)

- X14: income ($37500 - $49999, $62500 - $74999, $12500 - $24999, $75000 - $87499,
$50000 - $62499, $25000 - $37499, $100000 or More, $87500 - $99999, Less than $12500)
- X15: Car
- X16: Bar (never, less1, 1~3, gt8, nan4~8 (feature meaning: how many times do you go to a bar every month?))
- X17: CoffeeHouse (never, less1, 4~8, 1~3, gt8, nan (feature meaning: how many times do you go to a coffeehouse every month?))
- X18: CarryAway (n4~8, 1~3, gt8, less1, never (feature meaning: how many times do you get take-away food every month?))
- X19: ResturantLessThan20 (4~8, 1~3, less1, gt8, never (feature meaning: how many times do you go to a restaurant with an average expense per person of less than $20 every month?))
- X20: Restaurant20To50 (1~3, less1, never, gt8, 4~8, nan (feature meaning: how many times do you go to a restaurant with average expense per person of $20 - $50 every month?))
- X21: toCoupon_GEQ5min (0,1 (feature meaning: driving distance to the restaurant/bar for using the coupon
- X22: toCoupon_GEQ15min (0, 1 (feature meaning: driving distance to the restaurant/bar for using the coupon is greater than 15 minutes)
- X23: toCoupon_GEQ25min (0,1 (feature meaning: driving distance to the restaurant/bar for using the coupon is greater than 25 minutes))
- X24: direction_same (0, 1 (feature meaning: whether the restaurant/bar is in the same direction as your current destination))
- direction_opp (1, 0 (feature meaning: whether the restaurant/bar is in the same direction as your current destination))
- X26: Y The value 1 is considered as Yes and 0 as No (1, 0 (whether the coupon is accepted))

As the dataset has a lot of columns, to verify the correlations among the columns we have designed the correlation matrix as a heatmap.

Optimal Hyperparameters

- Hyperparameters for random forest:
{'n_estimators': 1000,
'min_samples_split': 2,
'min_samples_leaf': 1,
'max_features': 'sqrt',
'max_depth': 110,
'bootstrap': True}
- Hyperparameters for decision tree:
{'ccp_alpha': 0.0,
'class_weight': None,
'criterion': 'entropy',
'max_depth': 20,
'max_features': None,
'max_leaf_nodes': None,
'min_impurity_decrease': 0.0,
'min_impurity_split': None,
'min_samples_leaf': 50,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'random_state': 42,
'splitter': 'best'}
- Hyperparameters for logistic regression:
{'C': 1.0,
'class_weight': None,
'dual': False,
'fit_intercept': True,
'intercept_scaling': 1,
'l1_ratio': None,
'max_iter': 1000,
'multi_class': 'auto',
'n_jobs': None,
'penalty': 'l2',
'random_state': None,
'solver': 'lbfgs',

```
'tol': 0.0001,
'verbose': 0,
'warm_start': False}
```

Correlation HeatMap

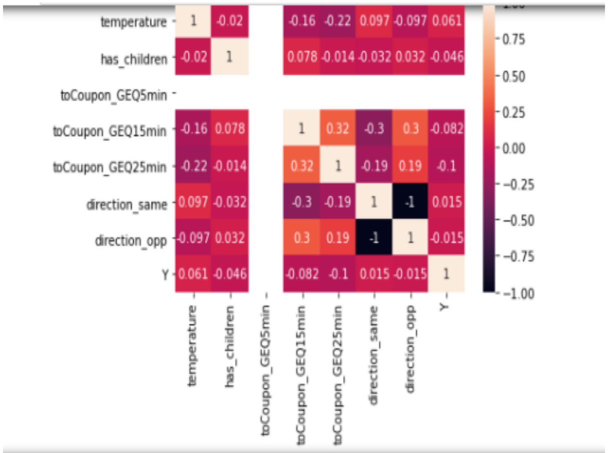


Figure 1.

This heatmap in Figure 1. depicts that the direction_same and direction_opp columns are both oppositely correlated with other. Therefore, the dataset will not be affected if one the column can be removed.

The variable transformation and the selection are done using Weight of Evidence (WOE) and Information Value (IV). It helps to handles the missing values, outliners and there is no need of dummy values to the variables. Moreover, the most of the transformation is based on logarithmic value of distributions.

$$WOE = \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

$$IV = \sum(\text{Event\%} - \text{Non Event\%}) * \ln\left(\frac{\text{Event\%}}{\text{Non Event\%}}\right)$$

or simply,

$$IV = \sum(\text{Event\%} - \text{Non Event\%}) * (\text{WOE})$$

Figure 2.

The data is divides into two groups; one is for training the model other to validate the model. The error rate is insensitive to the classification accuracy of models, as the accuracy rate calculated by the three different models are (67-75%). The accuracy performance of all the three algorithms is compared using ROC (Receiver Operating Characteristic Chart) curve chart. It helps to present the performance of a classification model at all classification thresholds.

- True Positive Rate
- False Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Figure 3.

3. Results

The ROC curve charts for the all three machine learning algorithms are shown below in Figures 4-7.

Random Forest ROC Curve

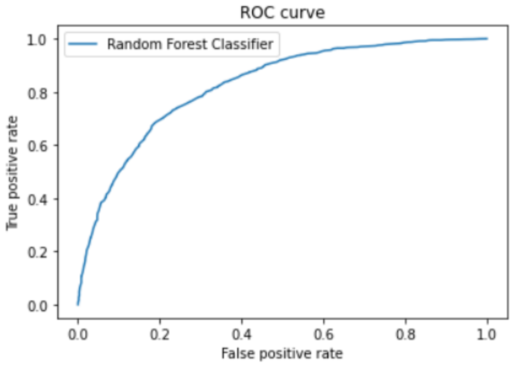


Figure 4.

Decision Tree ROC Curve

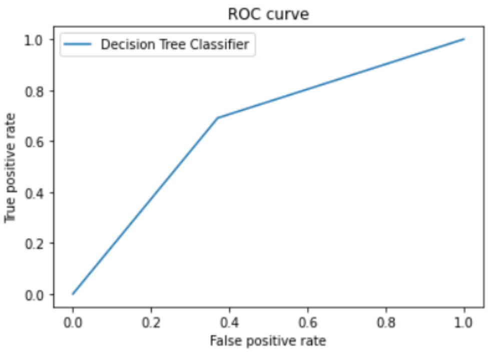


Figure 5.

Decision Tree Graph Visualization

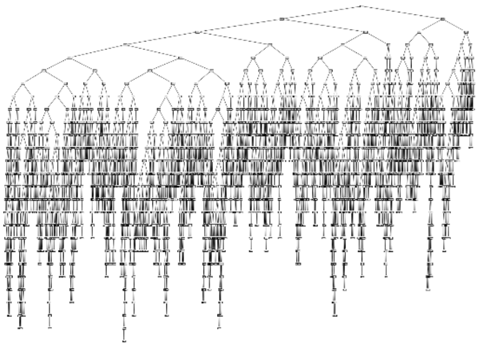


Figure 6.

Logistic Regression ROC Curve

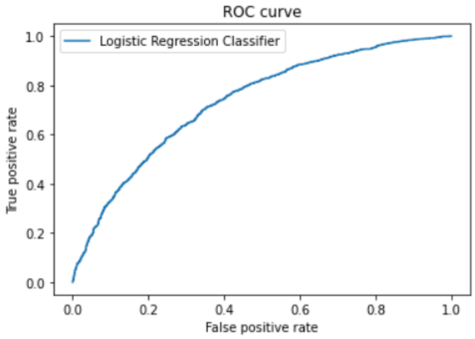


Figure 7.

Algorithms	**Random Forest**	**Decision Tree**	**Logistic Regression**
Accuracy Score	0.7491721854304636	0.6641832229580574	0.6862582781456954
ROC AUC (Area Under the Curve)	0.8248800269630124	0.6601024718156127	0.7335493157207967

Table1.

Confusion Matrix is used to evaluate the quality of a classifier on the data set. The diagonal values represent the number of points for which the predicted label is equal to the true label and the false label, mislabeled elements by the classifier are labelled by off-diagonal elements. The confusion matrix for all the three different machine learning algorithms are depicted in Figure 7-9.

Confusion matrix for Random Forest

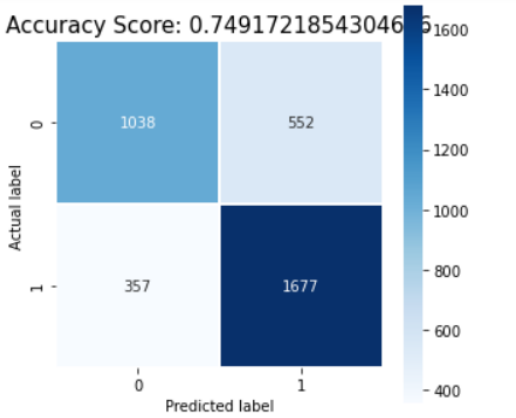


Figure 8.

Confusion matrix for Decision Tree

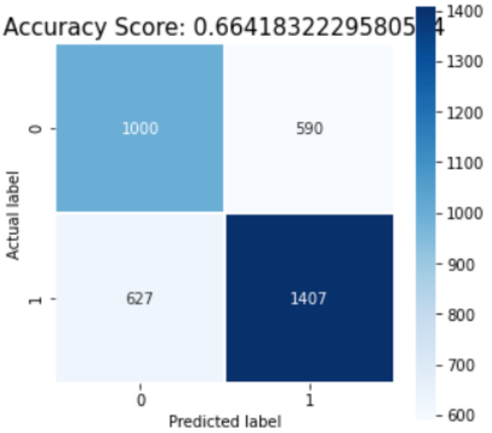


Figure 9.

Confusion matrix for Logistic Regression

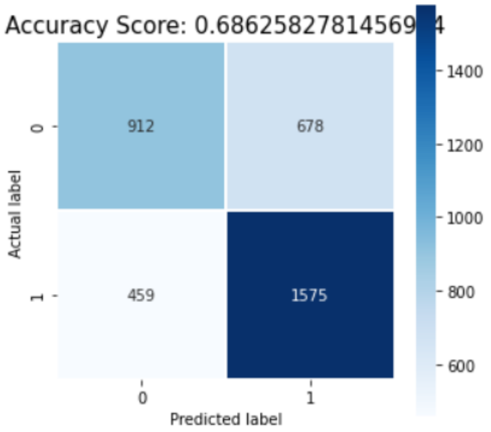


Figure 10.

The classification report visualizes the precision, recall and f1 score for the data model. The metrices are defined in the termed of True Positive and False Negatives. True positive is when the actual class and the estimated class both are true and whereas the false positive is when actual class is negative and estimated class is positive.

Precision: It is used to measure the classifier exactness and also the ratio of the true positives to the sum of true positive and true negatives.

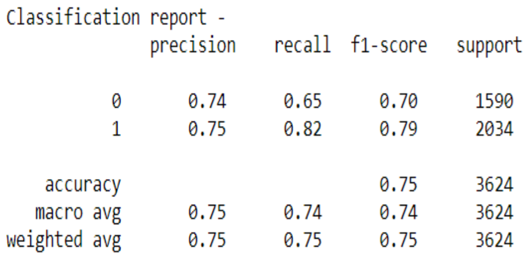
Recall: It is used to measure the classifier completeness and also ratio of true positives to the sum of true positives and false negatives.

F1 score: The weighted average of F1 should be used to compare classifier models.

Support: It is the number of actual occurrences of the class in the dataset.

The classification report for all the three different machine learning algorithms are depicted in Figure 10-12

Random Forest Classification Report



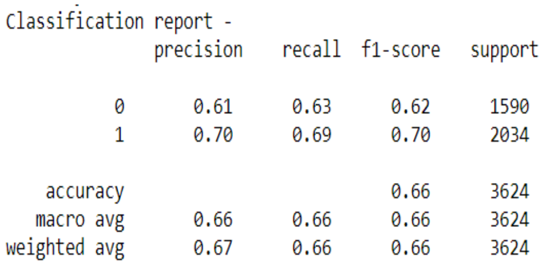
```
Classification report -
                  precision    recall  f1-score   support

           0       0.74      0.65      0.70      1590
           1       0.75      0.82      0.79      2034

    accuracy                           0.75      3624
   macro avg       0.75      0.74      0.74      3624
weighted avg       0.75      0.75      0.75      3624
```

Figure 11.

Decision Tree Classification Report



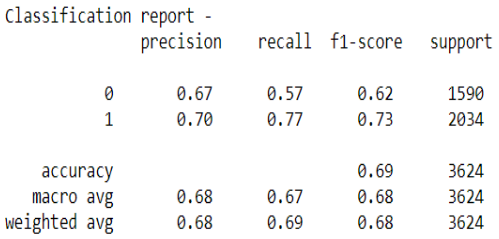
```
Classification report -
                  precision    recall  f1-score   support

           0       0.61      0.63      0.62      1590
           1       0.70      0.69      0.70      2034

    accuracy                           0.66      3624
   macro avg       0.66      0.66      0.66      3624
weighted avg       0.67      0.66      0.66      3624
```

Figure 12.

Logistic Regression Classification Report



```
Classification report -
                  precision    recall  f1-score   support

           0       0.67      0.57      0.62      1590
           1       0.70      0.77      0.73      2034

    accuracy                           0.69      3624
   macro avg       0.68      0.67      0.68      3624
weighted avg       0.68      0.69      0.68      3624
```

Figure 13.

CONCLUSION

This paper compares between the three machine learning algorithms and compares the performance in the form accuracy score. In the classification, the accuracy amounts the tree algorithms the results shows that there is a little difference between the accuracy values whereas the AUC values have also a little difference. The Random Forest perform the classification more accurately than the other two algorithms. Therefore, the Random Forest should be used to classify the coupon recommendation for the in-vehicle customers

REFERENCES

(1) T. Wang, C. Rudin, F. Velez-Doshi, Y. Liu, E. Klampfl and P. MacNeille, "Bayesian Rule Sets for Interpretable Classification," 2016 IEEE 16th International Conference on Data Mining (ICDM), 2016, pp. 1269-1274, doi: 10.1109/ICDM.2016.0171.

(2) M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," in IEEE Access, vol. 9, pp. 114381-114391, 2021, doi: 10.1109/ACCESS.2021.3104357.

(3) A. Smirnov and N. Shilov, "Intelligent driver support: Integration of coupon services into on-board infotainment systems," *2014 International Conference on Connected Vehicles and Expo (ICCVE)*, 2014, pp. 1043-1044, doi: 10.1109/ICCVE.2014.7297506

(4) Saeed Mirshekari, “Coupon Purchase Prediction” Competition in Kaggle, 2015.

(5) Cheng Yeh a, Che-hui Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients” pp. 2473-2480, 2009, doi: 10.1016/j.eswa.2007.12.020.

(6)https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

(7)https://www.scikit-yb.org/en/latest/api/classifier/classification_report.html

(8)https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc