**Name:** Garima Sharma

**Student ID:** 300343374

**Title:** Predict whether a person will accept the coupon recommended to him in different driving scenarios

1. **Introduction and discovery**

The dataset is taken from UCI, a Machine Learning Repository .It is a centre for machine learning and intelligent systems at UC Irvine. Because of research in field of machine learning and intelligent systems, these systems cop with the problem of developing the computer algorithms that can take and use the data in an intelligent way to solve a variety of real-world problems.

This data was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios including the destination, current time, weather, passenger, Time etc., and then ask the person whether he will accept the coupon if he is the driver. This data studies whether a person will accept the coupon recommended to him in different driving scenarios.

Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. This could include anything from conducting simple data validation and research to more subjective tasks like survey participation, content moderation, and more. Here, we want to predict, which driving scenarios are affecting the person to choose to accept or reject the coupon.

1. **Additional data preparation:**

There is no additional data preparation done.

Chart, bar chart

Description automatically generated

Out of 100%, nearly 57% belongs to positive class

Chart, bar chart

Description automatically generated

From above graph, the graph shows that people at time, specifically at 18 PM are highly likely to accept the coupons. They are less likely to accept the coupons when time is 14PM, 10AM.

Chart, bar chart

Description automatically generated

From here, we can see that we have higher counts when people do not have children.

1. **Model Implementation**:

According to my dataset, as my data contains categorical variables, I prefer to use Logistic Regression, Logistic Regression w/PCA, Decision Tree Classifier, Decision Tree Classifier w/PCA, Gradient Boosting Classifier, Gradient Boosting Classifier w/PCA,

Random Forest Classifier, Random Forest Classifier w/PCA, Ada Boost Classifier, Ada Boost Classifier w/PCA, LGBM, LGBM w/PCA, Cat Boost Classifier, Cat Boost Classifier w/PCA, KN Classifier, KN Classifier w/PCA, Support Vector Machine, Support Vector Machine w/PCA, MLP Classifier and MLP Classifier w/PCA. I scaled my data using Robust scaler and before scaling the features, I used the variance threshold feature selection technique to select the features. After that, different estimators are fit and GridSearchCV is also used with different grid parameters to improve the performance and also for hyper tuning. I fitted all the models using pipe and grid search cv then looked for the model, giving me the best result. I looked for the accuracy, recall, f1-score, and precision along with confusion matrix. Then I choose that model for further prediction analysis. At the end, best model would be used for analysis.

1. **Results Interpretation and Implication:**

Here are the results that I got:

Table

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidence

From here we can see, LGBM classifier perform well with 0.718188 i.e. with 72% of accuracy. Also, Precision is the ratio of correctly predicted positive observations to the total predicted positive observations and high precision relates to the low false positive rate. We have got 0.739201 precision which is not bad. Moreover, We have got recall of 0.790416 which is good for our model as it’s above 0.5. In addition to this, we have F1 score as 0.763951, which is good as our score is 76.4%.

Also, from the confusion matrix, we can see that we have 2156(787+1369) predictions that are predicted correctly. Whereas, we have 846(363+483) predictions that are not predicted correctly.

Chart, treemap chart

Description automatically generated

From the Classification report, we can see that

Table

Description automatically generated

From here, Performance for class 1 is better with regard to recall, f1-score and precision which is 79%, 76% and 74% respectively in our case. Also, our accuracy is 72%, which means our model is 72% accurate.

From ROC curve, we got:

Chart, line chart

Description automatically generated

The area under the ROC curve is a summary measure that essentially averages diagnostic accuracy across the spectrum of test values. That area under the curve is 71%, better the performance of the model at distinguishing between the positive and negative classes.

1. **Out-of-sample Predictions:**

Using my final model, which is LGBM, I performed predictions using new data and here is the sample of predicted probabilities.

**Text

Description automatically generated**

Here, it shows the probability of whether a person will accept the coupon recommended in different driving scenarios. The columns on left side, shows response as '0' and column on right shows response as '1'. In each case, such as [0.46551759 0.53448241], the probability on right side is higher as compared to the left column. So, in the case, the predicted probability for this one is 1.

Chart

Description automatically generated

From Confusion matrix, we can see that we have 462(179+283) predictions that are predicted correctly. Whereas, we have 138(42+96) predictions that are not predicted correctly.

Table

Description automatically generated

From here, Performance for class 1 is better with regard to recall and f1-score which is 87% and 80% respectively in our case. Also, our accuracy is 77%, which means our model is 77% accurate.

Chart, line chart

Description automatically generated

The area under the ROC curve is a summary measure that essentially averages diagnostic accuracy across the spectrum of test values. That area under the curve is 76%, better the performance of the model at distinguishing between the positive and negative classes.

1. **Concluding Remarks:**

The unseen dataset accuracy is greater than training accuracy which is 77%. This means that the model has generalized fine. Using GridSearchCV, we get accuracy/loss for every combination of hyper parameters, and we can choose the one with the best performance. In our case, it is LGBM classifier. As if our model is imbalanced then f1-score will take care of that. We can also try stratified sampling in future to improve the performance of the model.