In Vehicle Coupon Recommendation

—Data Science Capstone Project Presentation

Shixin Li
05/28/2021

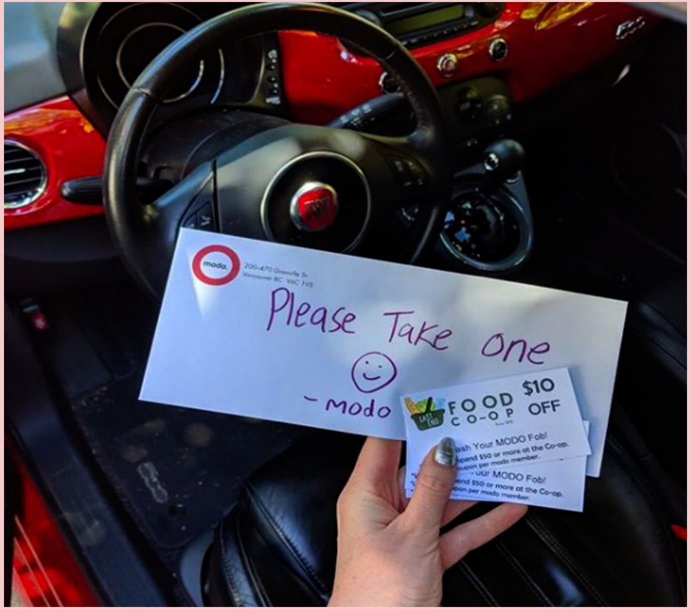
Contents

1. **Problem Statement**
2. **Data Information**
3. **Exploratory Data Analysis**
4. **Data Pre-processing**
5. **Model Selection**
6. **Conclusion**

Problem Statement

What is the problem?

- **In what a driving scenario, people would accept the recommended catering coupon?**



Who Care About this Problem?

- Catering Business

  * Restaurant

  * Coffee Shop

  * Bar

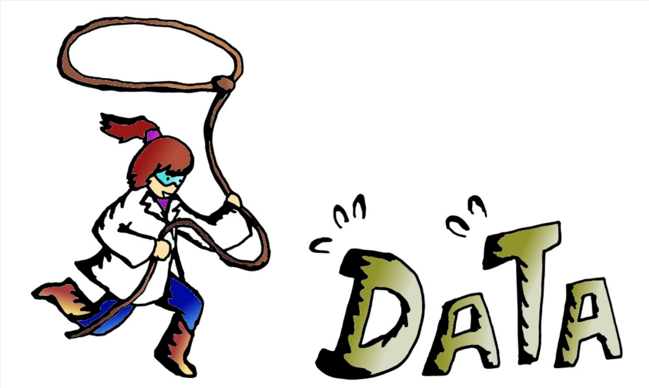
  ……

What to do?

- **Create supervised learning models to predict who would accept the recommended coupon in vehicle**

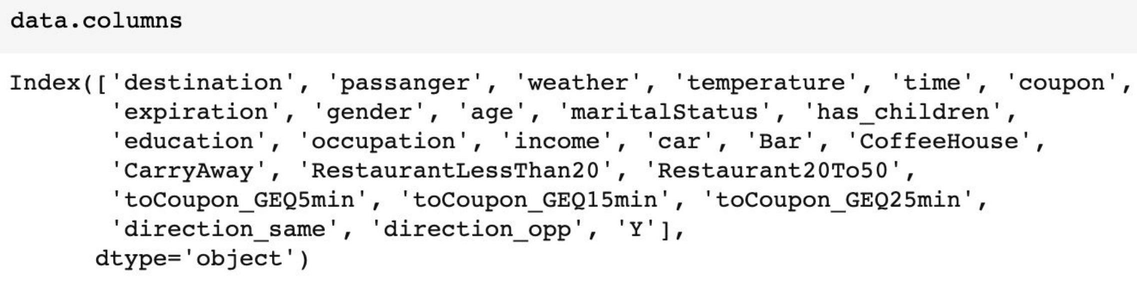
<u>Criteria for success:</u> Achieving at least 80% accuracy

Data Wrangling



Data Information

- **Collected via a survey on Amazon Mechanical Truck**
- **12684 observation, 26 columns with 1 column as label Y**
- **All features are categorical type**
- **Column name:**

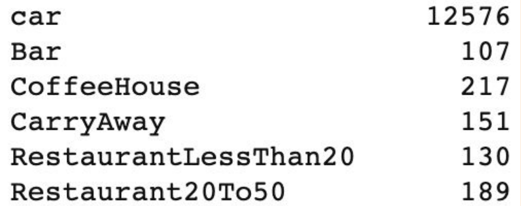


```
data.columns
```

```
Index(['destination', 'passanger', 'weather', 'temperature', 'time', 'coupon',
       'expiration', 'gender', 'age', 'maritalStatus', 'has_children',
       'education', 'occupation', 'income', 'car', 'Bar', 'CoffeeHouse',
       'CarryAway', 'RestaurantLessThan20', 'Restaurant20To50',
       'toCoupon_GEQ5min', 'toCoupon_GEQ15min', 'toCoupon_GEQ25min',
       'direction_same', 'direction_opp', 'Y'],
      dtype='object')
```

Data Cleaning

- **Missing Values:**



```
car                        12576
Bar                          107
CoffeeHouse                  217
CarryAway                    151
RestaurantLessThan20         130
Restaurant20To50             189
```

99.1% of values in the column ‘care’ are missing, **remove this column**

4.8% observations have missing values, **remove these rows**

- **Column with unique values:**

  * ToCouponEGQ5min

Remove this column

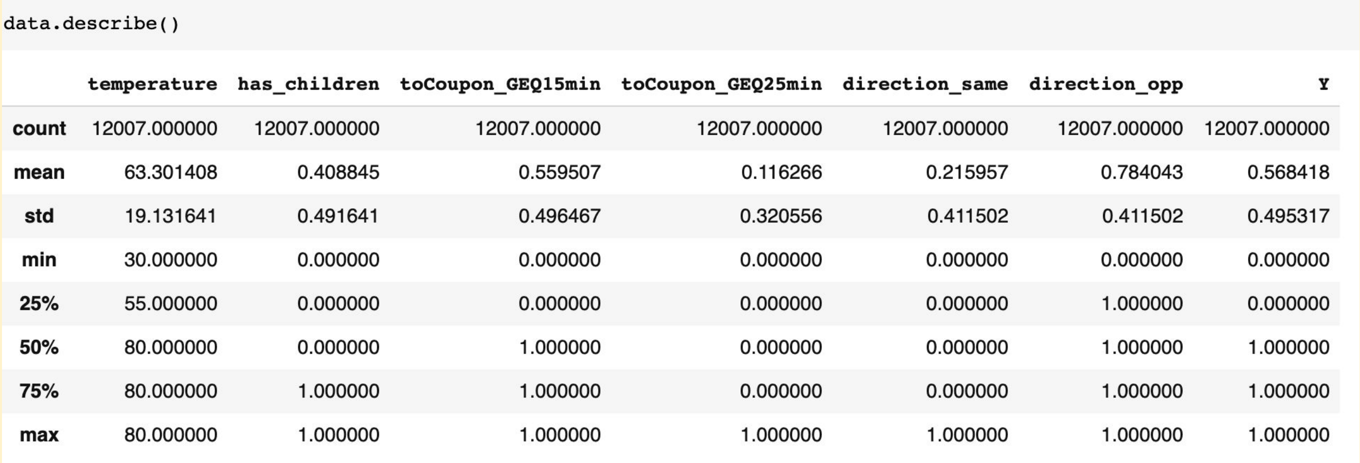
- 74 rows of duplicate rows

Remove these 74 rows

After this sept, 12007 rows and 24 columns are left in the data set.

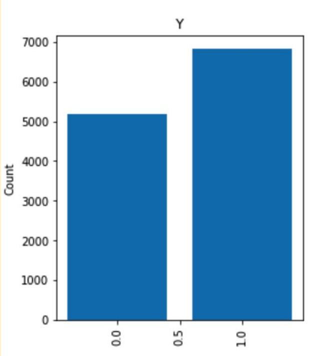
Data Exploratory Analysis

EDA — Statistical Data



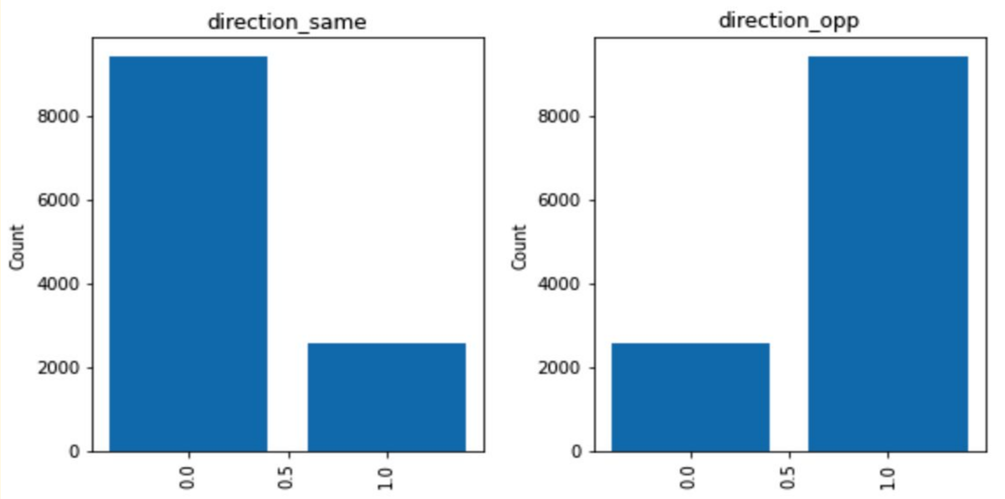
```
data.describe()
```

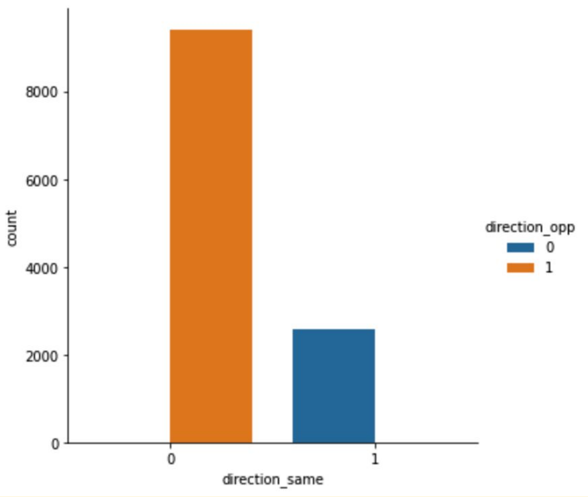
	temperature	has_children	toCoupon_GEQ15min	toCoupon_GEQ25min	direction_same	direction_opp	Y
count	12007.000000	12007.000000	12007.000000	12007.000000	12007.000000	12007.000000	12007.000000
mean	63.301408	0.408845	0.559507	0.116266	0.215957	0.784043	0.568418
std	19.131641	0.491641	0.496467	0.320556	0.411502	0.411502	0.495317
min	30.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	55.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
50%	80.000000	0.000000	1.000000	0.000000	0.000000	1.000000	1.000000
75%	80.000000	1.000000	1.000000	0.000000	0.000000	1.000000	1.000000
max	80.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000



- **For using the coupon, usually more than 25 mins to drive, not at the same direction as the destination, and the weather is higher than 55 Fahrenheit.**
- **56.7% of data are labeled 1 and 43.3% are labeled 0**

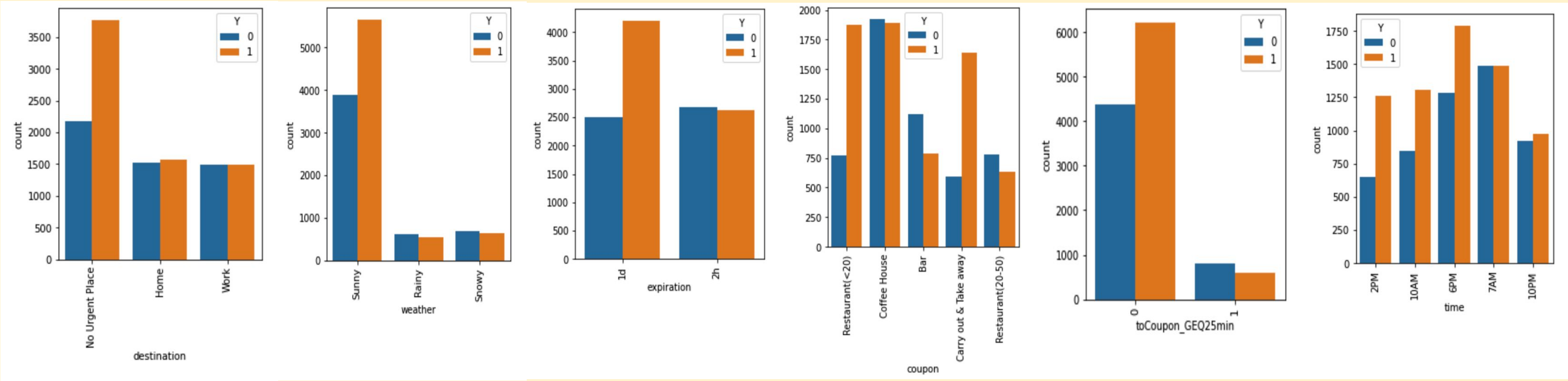
EDA—Data Visualization





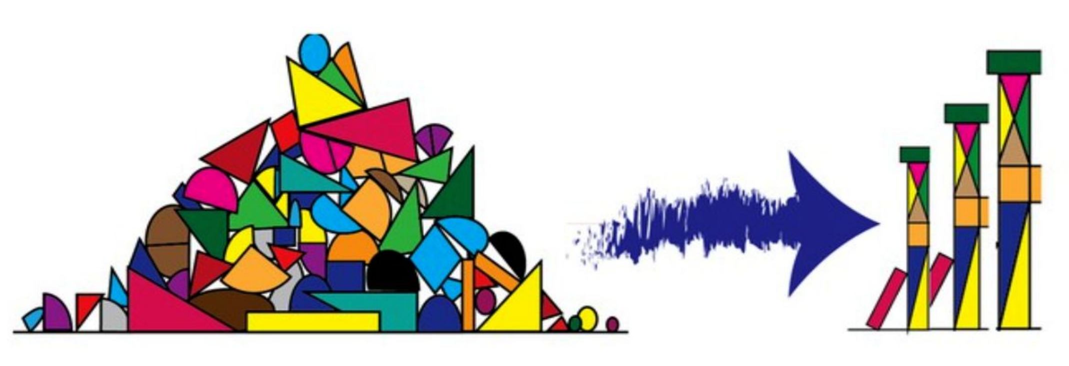
From the plot above, we can see that the feature ‘direction_same’ and ‘direction_opp’ are indicating the same fact, so we should remove one of them.

EDA—Data Visualization



People are more likely to use the coupon when they are going to not urgent place, in sunny day, the coupon will expiration sooner, the coupon for restaurant and take away, driving distance is less than 25 mins, in the afternoon.

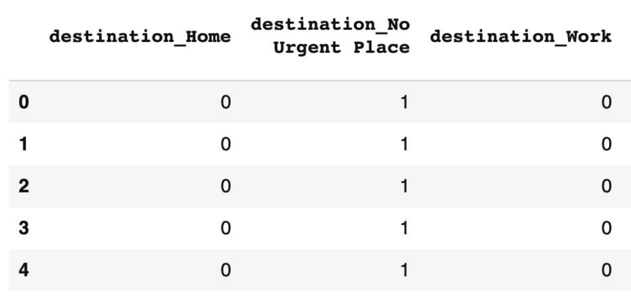
Data Pre-processing



Create Dummy Variables

After the process of EDA, the data set has 12007 rows and 24 columns.

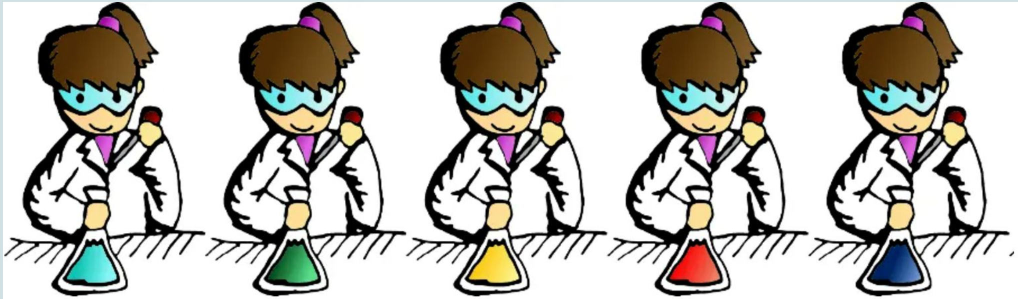
- **All features are categorical type, then transfer all of them to dummy variables**
- **After the transformation, the data set now has 113 columns**



	destination_Home	destination_No Urgent Place	destination_Work
0	0	1	0
1	0	1	0
2	0	1	0
3	0	1	0
4	0	1	0

Then split the data into training and test set with test size ratio 0.2, now the data is ready for training the model.

Model Selection



Model Training & Grid Search CV

Supervised Learning Models:

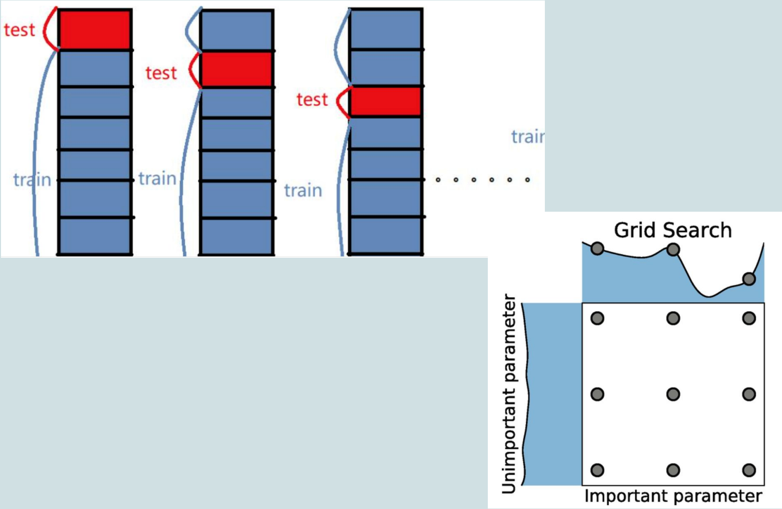
- Logistic Regression, Random Forest, K-nearest Neighbors, Support Vector Machine, Naive Bayes, Gradient Boosting

Hyperparameter Optimization：

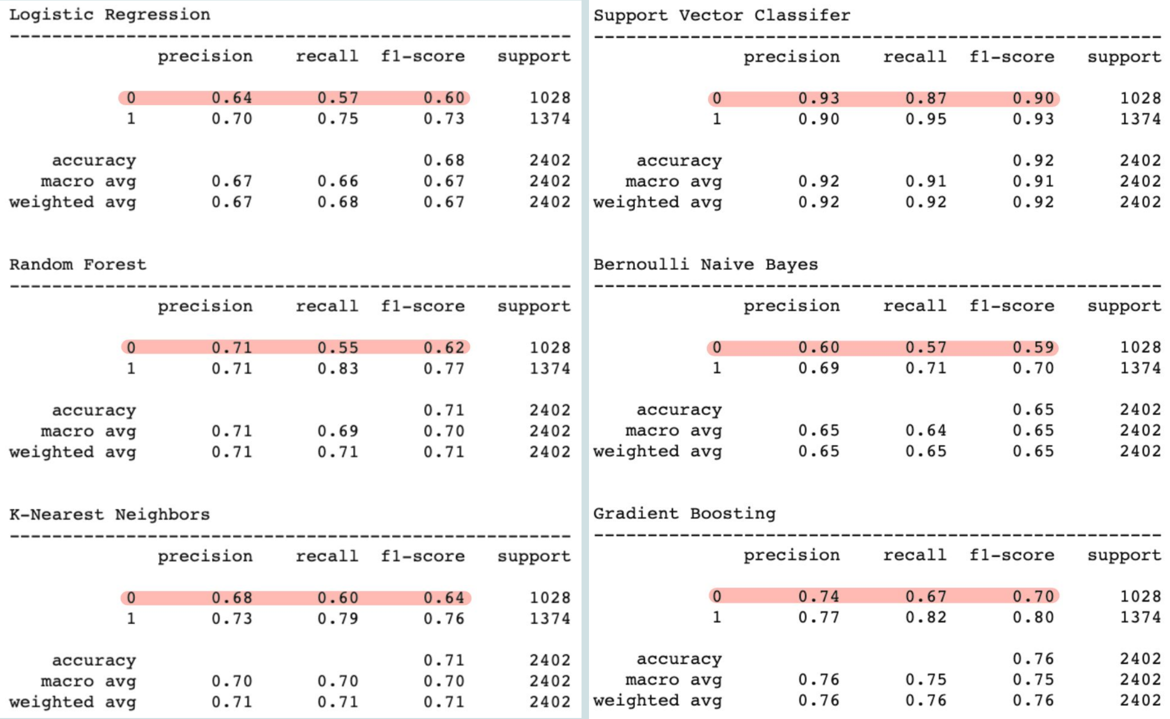
- Grid Search Cross Validation

Metric:

- F1-score



Result: F1-score and Accuracy

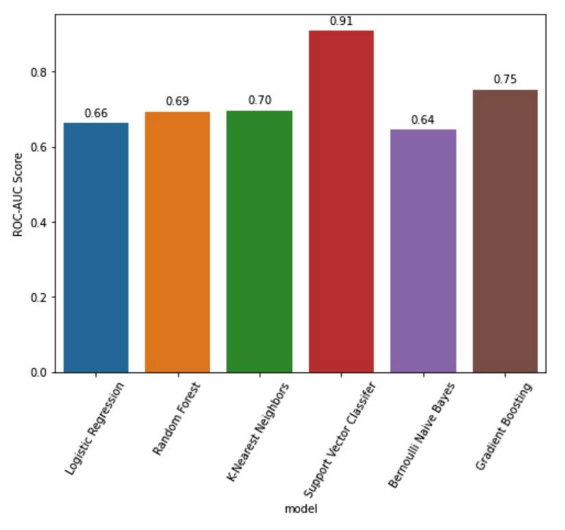


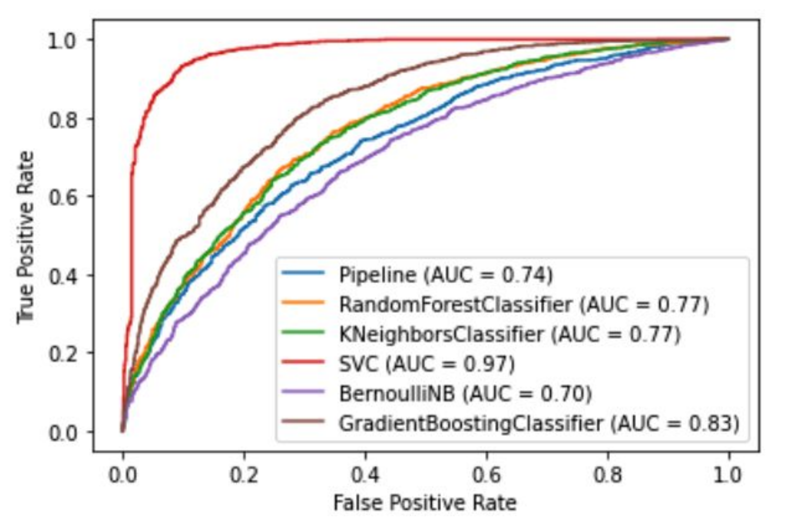
```
Logistic Regression
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.64      0.57      0.60      1028
           1       0.70      0.75      0.73      1374

    accuracy                           0.68      2402
   macro avg       0.67      0.66      0.67      2402
weighted avg       0.67      0.68      0.67      2402


Random Forest
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.71      0.55      0.62      1028
           1       0.71      0.83      0.77      1374

    accuracy                           0.71      2402
   macro avg       0.71      0.69      0.70      2402
weighted avg       0.71      0.71      0.71      2402


K-Nearest Neighbors
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.68      0.60      0.64      1028
           1       0.73      0.79      0.76      1374

    accuracy                           0.71      2402
   macro avg       0.70      0.70      0.70      2402
weighted avg       0.71      0.71      0.71      2402
```

```
Support Vector Classifer
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.93      0.87      0.90      1028
           1       0.90      0.95      0.93      1374

    accuracy                           0.92      2402
   macro avg       0.92      0.91      0.91      2402
weighted avg       0.92      0.92      0.92      2402


Bernoulli Naive Bayes
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.60      0.57      0.59      1028
           1       0.69      0.71      0.70      1374

    accuracy                           0.65      2402
   macro avg       0.65      0.64      0.65      2402
weighted avg       0.65      0.65      0.65      2402


Gradient Boosting
-------------------------------------------------------
              precision    recall  f1-score   support

           0       0.74      0.67      0.70      1028
           1       0.77      0.82      0.80      1374

    accuracy                           0.76      2402
   macro avg       0.76      0.75      0.75      2402
weighted avg       0.76      0.76      0.76      2402
```

- **All model have lower f1-score on negative than positive.**
- **Naive Bayes model has the lowest f1-score and accuracy**
- **Support Vector Classifier has the highest f1-score, 0.915, which is at least 0.2 higher than other models**
- **Support Vector Classifier has the highest accuracy, 92%.**

Result: ROC-AUC Score & AUC Value





Support Vector Classifier has highest ROC-AUC Score and AUC Value, 0.91 and 0.97, which are much better than other models.

Conclusion

Conclusions :

- Base on the f1-score, ROC-AUC score and ROC-AUC value, Support Vector Classifier(SVC) is the best model for the In Vehicle Coupon Recommendation Data Set
- All features, total 112 dummy variables, are applied to the SVC
- The accuracy and f1-score of SVC are 0.915 and 92%
- Only the SVC achieved the Criteria for success, a accuracy of 80%

Ideas for Further Research

- Consider how much people usually spent in meal or drinking
- Where people are living in, urban or rural areas
- What kind of food, drink or alcohol they prefer
- A similar research can also applied on the big supermarket, which can help them allocate the commodity and arrange the shops better

The End!!!!!

Thank you for watching!!!!