**Project: In-Vehicle Coupon Recommendation**

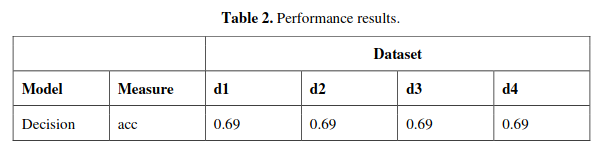
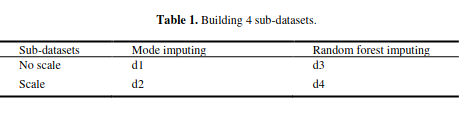
**“We have some catching up to do in the area of machine learning and artificial intelligence.” ~Klaus Froehlich**

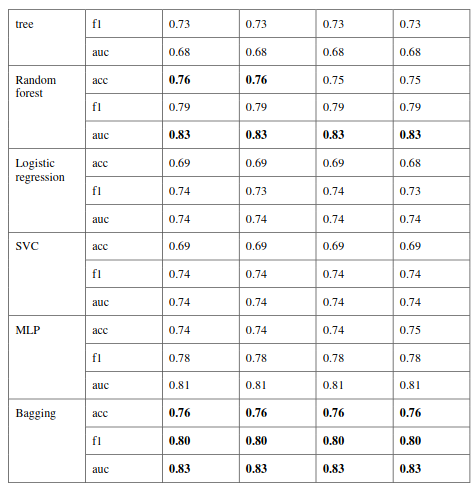
**Introduction**

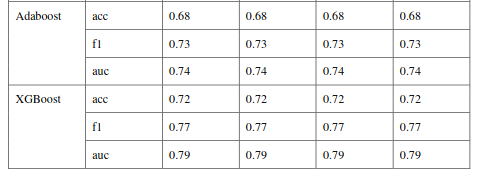
To successfully grows a business, understanding customer behavior is essential [1]. In recent years machine learning researchers have worked on customer behavior analysis [1]. Many big tycoon companies such as Amazon, Google, and Facebook invest time, effort, and money in machine learning only to understand the behavior of users so they can make business decisions[2].

In this project we have worked on “[In-Vehicle Coupon Recommendation](https://www.kaggle.com/datasets/mathurinache/invehicle-coupon-recommendation)”, training a machine learning model that predicts customer behavior that it will accept a coupon of a nearby location(restaurant or coffee house, or bar) based on user, contextual and coupon information.

There is a research paper that uses the following technique.







In this we applied the following techniques:

| **Model Name** | **Accuracy** | **AUC** | **F1** |
| --- | --- | --- | --- |
| KNN | 0.70 | 0.74 | 0.70 |
| Naive Bayes | 0.68 | 073 | 0.68 |
| Logistic Regression | 0.71 | 0.76 | 0.71 |
| SVM | 0.73 | 0.778 | 0.73 |
| Random Forest | 0.68 | 0.73 | 0.68 |
| CatBoost | 0.72 | 0.777 | 0.72 |
| ANN | 0.71 | 0.76 | 0.71 |

To keep in view above we apply hyperparameter tuning on SVM model and deploy it on a Flask.

**Problem statement**

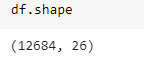
Understanding customer behavior is essential for business decision-making. Business decision-making is a very critical task because it involves the money and effort of companies, and any misinformation can lead to a big problem.

The dataset that we use in this project has problems such as duplicates, missing values, and many unnecessary data(noise). There is a need for pre-processing before moving toward modeling.

**Proposed Methodology**

* Explored the dataset and found problems in it.
  + Features and shape:

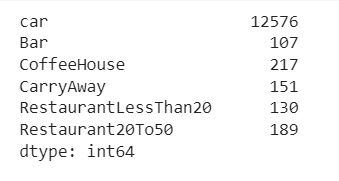




* + There is a feature named toCoupon\_GEQ5min that has a constant value.

toCoupon\_GEQ5min : [1]

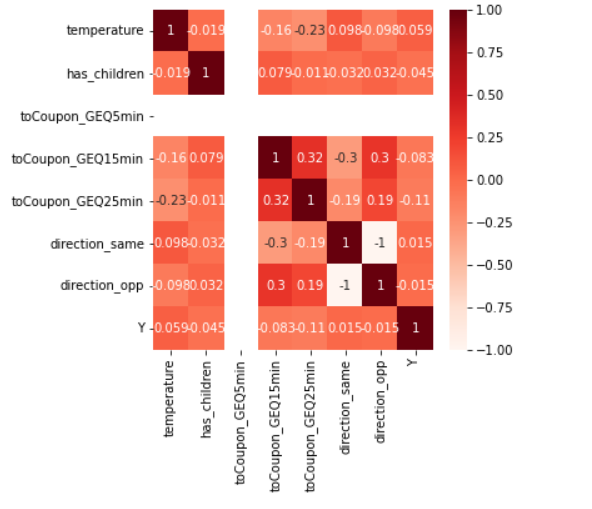
* + Missing values:



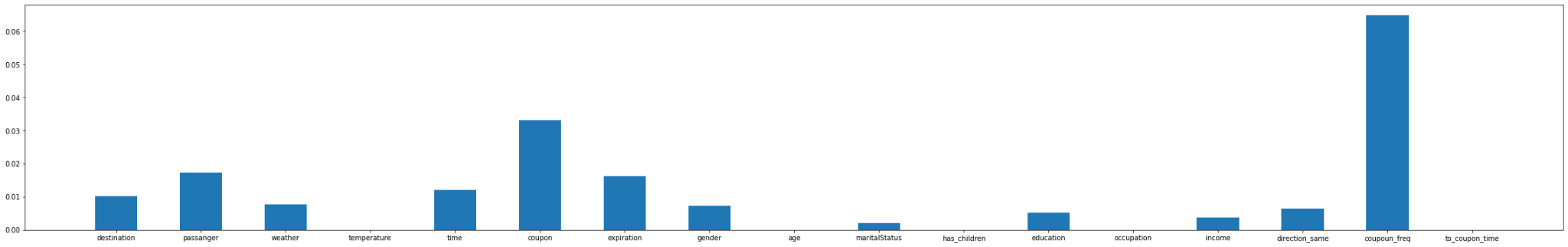
* + Duplicate data:



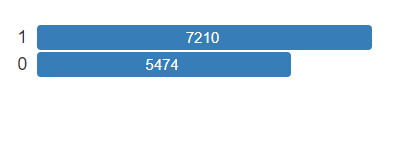
* + Correlation between features:



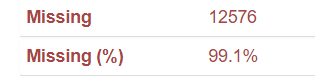
* + Most of the columns in the dataset are categorical and some of them are unnecessary that contribute nothing to prediction:



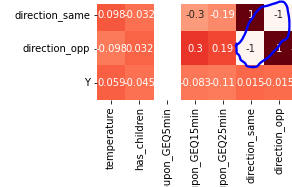
* + Class Imbalance:



* Applied pre-processing to solve the above-mentioned problems.
  + Drop duplicate data
  + Drop feature named car, because it is 99% null



* + Fill mode of features for filling missing values, because all the features we have are categorical.
  + Drop the feature named direction\_opp, because it is highly correlated with the feature named direction\_same



* + Extract a new feature named coupon\_freq from Bar, CoffeeHouse, CarryAway, RestaurantLessThan20, and Restaurant20To50 because we want only the feature that name is provided in the coupon column.



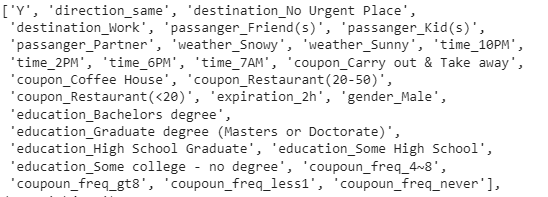
* + Extract a new feature named to\_coupon\_time from toCoupon\_GEQ15min and toCoupon\_GEQ25min, because these both are represented time so can merge them. 0 represents time is less than 15 minutes, 1 represents time is greater than or equal to 15 minutes but less than 25 minutes, and 2 represents time is greater than or equal to 25 minutes.



* + Find mutual information of features and select those features that have mutual information scores greater than 0.005.



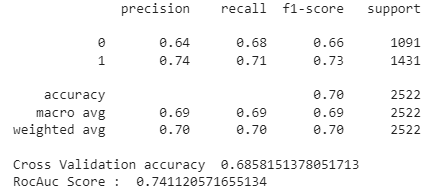
* + Convert categorical columns into dummy variables.



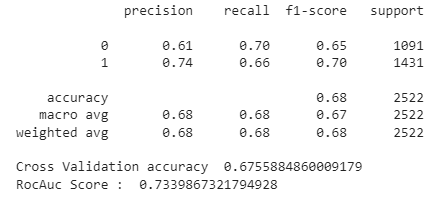
* + Apply smote to solve the class imbalance problem.
* Split data into three parts (train, test, and validation)



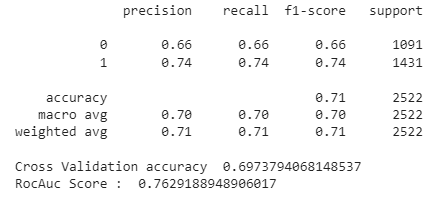
* Apply different classifiers and find the AUCROC score.
  + KNN



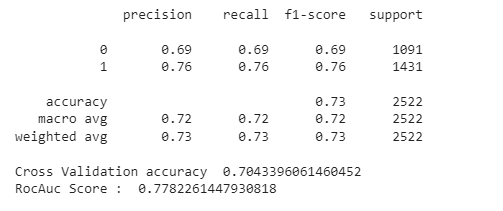
* + Naive Bayes



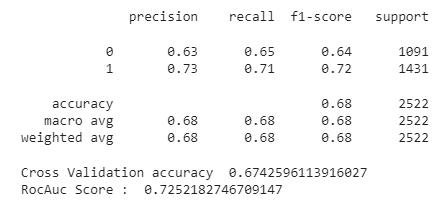
* + Logistic regression



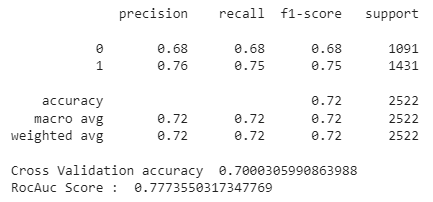
* + SVM



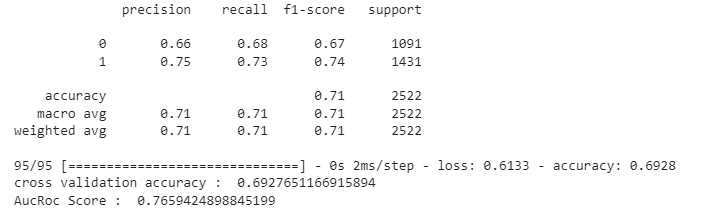
* + Random Forest



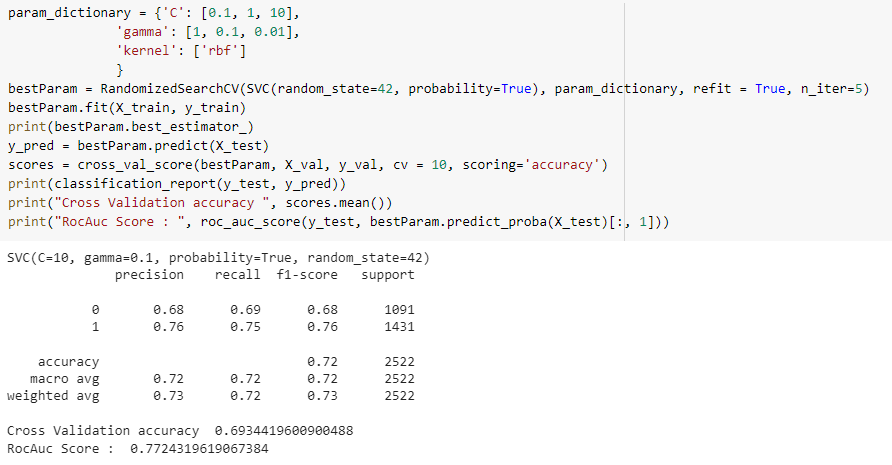
* + CatBoost



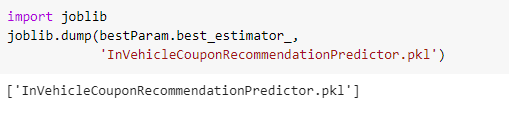
* + ANN



* Select the SVM classifier for the next steps because it has a better AUCROC score and accuracy.
* Tune hyperparameters of SVM using RandmizedSearchCV



* Save the model using Joblib and deploy it on the flask.





**Dataset Discussion**

This data was collected via a survey on Amazon Mechanical Turk. The survey describes different driving scenarios including the destination, current time, weather, passenger, etc., and then asks the person whether he will accept the coupon.

Basic information about Dataset:

| Name | In-Vehicle Coupon Recommendation |
| --- | --- |
| Problem | Binary Classification |
| no: instances | 12684 |
| no: features | 26 |
| duplicates | 74 |
| missing | yes |
| Classes frequency | 1: 7210, 0: 5474 |
| Year | 2017 |

There are 26 attributes.

**destination**

No Urgent Place 6283

Home 3237

Work 3164

**passenger**

Alone 7305

Friend(s) 3298

Partner 1075

Kid(s) 1006

**weather**

Sunny 10069

Snowy 1405

Rainy 1210

**temperature**

80 6528

55 3840

30 2316

**time**

6PM 3230

7AM 3164

10AM 2275

2PM 2009

10PM 2006

**coupon**

Coffee House 3996

Restaurant(<20) 2786

Carry out & Take away 2393

Bar 2017

Restaurant(20-50) 1492

**expiration**

1d 7091

2h 5593

**gender**

Female 6511

Male 6173

**age**

21 2653

26 2559

31 2039

50plus 1788

36 1319

41 1093

46 686

below21 547

**maritalStatus**

Married partner 5100

Single 4752

Unmarried partner 2186

Divorced 516

Widowed 130

**has\_children**

0 7431

1 5253

**occupation**

Some college - no degree 4351

Bachelors degree 4335

Graduate degree (Masters or Doctorate) 1852

Associates degree 1153

High School Graduate 905

Some High School 88

Name: education, dtype: int64

Unemployed 1870

Student 1584

Computer & Mathematical 1408

Sales & Related 1093

Education&Training&Library 943

Management 838

Office & Administrative Support 639

Arts Design Entertainment Sports & Media 629

Business & Financial 544

Retired 495

Food Preparation & Serving Related 298

Healthcare Practitioners & Technical 244

Healthcare Support 242

Community & Social Services 241

Legal 219

Transportation & Material Moving 218

Architecture & Engineering 175

Personal Care & Service 175

Protective Service 175

Life Physical Social Science 170

Construction & Extraction 154

Installation Maintenance & Repair 133

Production Occupations 110

Building & Grounds Cleaning & Maintenance 44

Farming Fishing & Forestry 43

**income**

$25000 - $37499 2013

$12500 - $24999 1831

$37500 - $49999 1805

$100000 or More 1736

$50000 - $62499 1659

Less than $12500 1042

$87500 - $99999 895

$75000 - $87499 857

$62500 - $74999 846

**car**

Scooter and motorcycle 22

Mazda5 22

do not drive 22

crossover 21

Car that is too old to install Onstar :D 21

**Bar**

never 5197

less1 3482

1~3 2473

4~8 1076

gt8 349

**CoffeeHouse**

less1 3385

1~3 3225

never 2962

4~8 1784

gt8 1111

**CarryAway**

1~3 4672

4~8 4258

less1 1856

gt8 1594

never 153

**RestaurantLessThan20**

1~3 5376

4~8 3580

less1 2093

gt8 1285

never 220

**Restaurant20To50**

less1 6077

1~3 3290

never 2136

4~8 728

gt8 264

**toCoupon\_GEQ5min**

1 12684

**toCoupon\_GEQ15min**

1 7122

0 5562

**toCoupon\_GEQ25min**

0 11173

1 1511

**direction\_same**

0 9960

1 2724

**direction\_opp**

1 9960

0 2724

**Y**

1 7210

0 5474

**Major Outcomes**

* Solve problems of the dataset, apply the feature extraction, and selection.
* Applying different classifiers and selecting the classifier based on the AUCROC score.
* Saving the model and deploying it on the flask.

**Project Timeline**

| Week | Task |
| --- | --- |
| 1 | Data Exploration |
| 2 | Preprocessing and Modeling |
| 3 | Hyperparameter Tuning, Deployment |
| 4 | Testing |

**Conclusion**

For taking business decisions understanding customer behavior is very important. Taking business decisions is a very critical task. In recent years machine learning emerge as a tool for understanding customer behavior and taking business decisions. In this project, we work on a recommendation system that predicts customer behavior that it will accept a coupon of a nearby location such as a restaurant or coffee house or bar, etc based on its context and user and coupon information.

**References**

1. Quynh, T. D., & Dung, H. T. T. Prediction of Customer Behavior using Machine Learning: A Case Study. In *Proceedings of the 2nd International Conference on Human-centered Artificial Intelligence (Computing4Human 2021). CEUR Workshop Proceedings, Da Nang, Vietnam (Oct 2021)*.
2. <https://spd.group/artificial-intelligence/ai-for-customer-behavior-analysis/>
3. Çelik, E., & Omurca, S. İ. Comparative Analysis of Offline Recommendation Systems with Machine Learning Algorithms. *PROCEEDINGS BOOK*.
4. Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., & MacNeille, P. (2017). A bayesian framework for learning rule sets for interpretable classification. *The Journal of Machine Learning Research*, *18*(1), 2357-2393.
5. https://medium.com/@niralidedaniya/in-vehicle-coupon-recommendation-a-machine-learning-classification-case-study-df67e7835703

**Abbreviations**

SVM: Support Vector Machine

ANN: Artificial Neural Network

KNN: KNearest Neighbour

AUCROC: Area under the ROC Curve

ROC: receiver operating characteristic curve