PROJECT REPORT
(PROJECT TERM JANUARY - MAY 2024)

COUPOUN PURCHASE PREDICTION

SUBMITTED BY

B SAI ASHISH
12323855

NIRBHAY KUMAR PANDEY
12302382

BODANAPATI VINAY
12312952

SECTION K22SA

COURSE CODE INT254

UNDER THE GUIDANCE OF Dr.DHANPRATAP SINGH
SCHOOL OF COMPUTER SCIENCE ENGINEERING

LOVELY PROFESSIONAL UNIVERSITY

DECLERATION

I hereby declare that the project work entitled “Coupoun Purchase Prediction” is an authentic record of my own work carried out as requirements of Project for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara, under the guidance of Dr Dhanpratap Singh, during JANUARY to MAY 2024. All the information furnished in this project report is based on my own intensive work and is genuine.

B SAI ASHISH
DATE:
12323855
05/04/2024

CERTIFICATE

This is to certify that the declaration statement made by this student is correct to the best of my knowledge and belief. He has completed this Project under my guidance and Supervision. The present work is the result of his original investigation, effort and study. No part of the work has ever been submitted for any other degree at any University. The Project is fit for the submission and partial fulfillment of the conditions for the award of B. Tech degree in Computer Science and Engineering from Lovely Professional University, Phagwara.

Dr Dhanpratap Singh
School of Computer Science and Engineering, Lovely Professional University,
Phagwara, Punjab
Date:
05/04/2024

ACKNOWLEDGEMENT

It is with my immense gratitude that I acknowledge the support and help of my Professor, Dr Dhanpratap Singh, who has always encouraged me into this research. Without his continuous guidance and persistent help, this project would not have been a success for me. I am grateful to the Lovely Professional University, Punjab and the department of Computer Science without which this project would have not been an achievement. I also thank my family and friends, for their endless love and support throughout my life.

TABLE OF CONTENTS

ABSTRACT

This study focuses on predicting coupon purchases using machine learning algorithms. The dataset used in this research consists of various features such as customer demographics, purchase history, coupon details, and time-related information. The goal is to develop a predictive model that can accurately forecast whether a customer will purchase a coupon or not based on these features.

To achieve this, several machine learning techniques such as logistic regression, random forest, and gradient boosting are employed and compared for their predictive performance. Feature engineering and selection play a crucial role in enhancing the model's accuracy and generalization capabilities.

The results of the study indicate that the random forest algorithm outperforms the other models in terms of predictive accuracy. By leveraging customer data and coupon information, businesses can better understand consumer behavior and tailor their marketing strategies to increase coupon purchases.

Overall, this research provides valuable insights into predicting coupon purchases and demonstrates the potential of machine learning in optimizing marketing efforts and enhancing customer engagement.

INTRODUCTION

- Coupon purchase prediction has become a crucial aspect of marketing strategies in today's competitive business landscape. With the rise of e-commerce and digital advertising, companies are constantly seeking ways to maximize their coupon effectiveness and increase customer engagement. By accurately predicting whether a customer is likely to purchase a coupon, businesses can target their marketing efforts more effectively and improve their return on investment.

- In this study, we develop into the realm of machine learning and data analytic to develop a predictive model for coupon purchases. By analyzing customer demographics, purchase history, coupon details, and other relevant features, we aim to create a model that can forecast coupon purchases with high accuracy.

- The ability to predict coupon purchases not only benefits businesses by optimizing their marketing campaigns but also enhances the overall customer experience. By offering personalized discounts and promotions to customers who are more likely to make a purchase, companies can foster customer loyalty and drive sales growth.

- Through this research, we aim to shed light on the importance of coupon purchase prediction in modern marketing strategies and demonstrate the potential impact of data-driven approaches in improving business outcomes However, it is not easy to handle large-scale data and extract features of consumers, companies, digital coupons and so forth.
- Large amount of data and complexity of behavior analysis make traditional methods in Econometric ineffective. Fortunately, the rising and development of data science in recent years provides a solution to this problem.
- Machine learning is a new way to analyze data and performs well in prediction . In this paper, the XGBoost algorithm in machine learning is used to establish a novel Digital Coupon Use Prediction Model.

RELATED WORDS

The study of coupon purchase prediction involves using machine learning techniques to analyze data related to customer behavior, purchasing patterns, and coupon usage in order to make predictions about future purchases. This area of study aims to understand and anticipate how customers are likely to use coupons and make purchases, which can be valuable for businesses in optimizing their marketing strategies and offers.

Using past purchase and browsing behavior, this competition asks you to predict which coupons a customer will buy in a given period of time. The resulting models will be used to improve Ponpare's recommendation system, so they can make sure their customers don't miss out on their next favorite thing.

Mean Average Precision @ 10 (MAP@10): MAP@10

$$MAP@10 = \frac{1}{|U|} \sum_{u=1}^{|U|} \frac{1}{\min(m, 10)} \sum_{k=1}^{\min(n,10)} P(k)$$

where I |U| is a number of users I m is the number of purchased coupons for the given user I n is the number of predicted coupons I P(k) is the precision at cutoff k:

$$P(k) = \frac{N \text{ correct coupons from the first } k \text{ predicted}}{k},$$

if k-th coupon is predicted correctly, 0 - otherwise.

CROSS VALIDATION

Problem: possible pairs user-coupon in train: ⇡ 360 000 000 I

- To decrease the size of train, for each week:
  - I take coupons with at least one purchase
  - I take users with at least one purchase
  - It gives ⇡ 600 000 pairs for each week, or ⇡ 30 000 000 pairs for the whole train
  - I use last few weeks to predict test week (we used last 5 weeks)

- Validation set: pairs for the last week
- To match CV and LB score: use multiplier (MAP@10 = 0 for users without purchases)

XGBOOST with **rank:map** objective and **map@10** evaluation metric

Place	Team	Leaderboard score
1	Herra Huu	0.009973
2	Halla Yang	0.009848
3	threecourse	0.009484
...	...	...
20	**Dmitry and Leustagos**	**0.007642**

Dataset Description

In this segment, we will delve into some of the best practices and methodologies for coupon data collection and preparation:

1. **Data Quality Assessment**: Before employing coupon data for machine learning tasks, it is essential to evaluate its quality and address any issues such as missing values, outliers, duplicates, errors, inconsistencies, or biases. Various techniques and tools can be utilized for data quality assessment, including descriptive statistics, data profiling, data visualization, data validation, and data cleansing. For instance, tools like pandas in Python can help explore and manipulate coupon data in tabular form, while visualization libraries like matplotlib or seaborn can aid in examining data distribution, correlation, and outliers through histograms, boxplots, scatterplots, or heatmaps.

2. **Data Integration and Transformation**: Coupon data from diverse sources may have distinct schemas, formats, units, scales, or granularity levels. To leverage coupon data for machine learning purposes, it is necessary to integrate and transform the data into a unified and compatible format and structure. Techniques such as data mapping, merging, aggregation, normalization, encoding, scaling, and imputation can be applied for data integration and transformation. For example, SQL or pandas can be used to combine coupon data from various tables or files, sklearn in Python can assist in encoding categorical data, and methods like min-max scaling or standardization can normalize numerical data.

3. **Data Enrichment and Feature Engineering**: Coupon data might lack certain relevant information crucial for machine learning tasks, such as customer demographics, preferences, behavior, or feedback. To optimize coupon data for machine learning applications, it is imperative to enrich and engineer the data by introducing additional and derived features that capture the data's characteristics, patterns, and relationships. Techniques like data augmentation, scraping, extraction,

selection, and creation can be employed for data enrichment and feature engineering. For instance, web scraping or APIs can be used to gather external coupon data, regex or NLP techniques can extract keywords or sentiments, and dimensionality reduction methods like PCA or LDA can be utilized. Additionally, domain expertise and business logic can aid in creating new features like coupon redemption rate or profitability from the coupon data.

- ✧ Related: association, correlation, connection
- ✧ Words: terms, vocabulary, language
- ✧ Coupon: voucher, discount, code
- ✧ Purchase: buy, acquire, obtain
- ✧ Prediction: forecast, estimation, anticipation

Resampling to Account for COUPOUN

Resampling techniques can be used to account for imbalanced data when dealing with coupon purchase prediction. By resampling the data, such as through oversampling the minority class or undersampling the majority class, the machine learning model can be trained on a more balanced dataset. This can help improve the model's ability to accurately predict coupon redemptions and purchases, especially when the occurrence of coupon redemptions is rare compared to non-redemptions.

In order to limit the extent to which our models will study the differences between 'Before News' and Reuters, we force the distribution to cover a limited range by negligently taking samples of the largest source n small n....
We selected 500 n-max articles for use as it seemed prudent, though not legally supported. Nor do we downgrade low frequency sources to maintain a certain range of resources. Correct n-max (or potential n-min) is a research question that is of interest to itself. Additional research methods may consider changing this number to get a better result if you are facing the same corpus difficulty. We note that before and after

TABLE II: Irregular and random model performance everywhere metrics.

survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
1	1	Cumings, I	female	38	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen,	female	26	0	0	STON/O2.	7.925		S
1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
0	3	Moran, M	male		0	0	330877	8.4583		Q
0	1	McCarthy,	male	54	0	0	17463	51.8625	E46	S
0	3	Palsson, M	male	2	3	1	349909	21.075		S
1	3	Johnson, I	female	27	0	2	347742	11.1333		S
1	2	Nasser, M	female	14	1	0	237736	30.0708		C
1	3	Sandstron	female	4	1	1	PP 9549	16.7	G6	S

FEATURE GENERATION

Feature generation for coupon purchase prediction involves creating informative input variables (features) from the available data that can be used to train a machine learning model to predict whether a customer will redeem a coupon and make a purchase. Some potential features for this task could include:

1. Customer demographics (age, gender, location)
2. Purchase history (previous purchases, frequency of purchases)
3. Coupon usage history (previous redemptions, types of coupons used)
4. Time-related features (day of the week, time of day)
5. Interaction features between customer behavior and coupon usage
6. Sentiment analysis of customer reviews or feedback
7. External factors (seasonality, economic indicators)

These features can be engineered from raw data and used to build a predictive model that can forecast coupon redemptions and purchases.

PRE-PROCESSING

Pre-processing for coupon purchase prediction involves several steps to prepare the data for machine learning model training. Some common pre-processing steps include:

1. Data Cleaning: Handling missing values, correcting inconsistencies, and removing irrelevant information.

2. Feature Scaling: Scaling numerical features to a similar range to prevent certain features from dominating the model.

3. Feature Encoding: Converting categorical variables into a numerical format that can be used by machine learning algorithms.

4. Feature Selection: Identifying and selecting the most relevant features to improve model performance and reduce dimensionality.

5. Data Splitting: Dividing the dataset into training, validation, and testing sets to evaluate the model's performance.

6. Handling Imbalanced Data: Addressing class imbalances by using techniques such as oversampling, undersampling, or generating synthetic samples.

By performing these pre-processing steps, the data can be better suited for training machine learning models to predict coupon redemptions and purchases accurately.

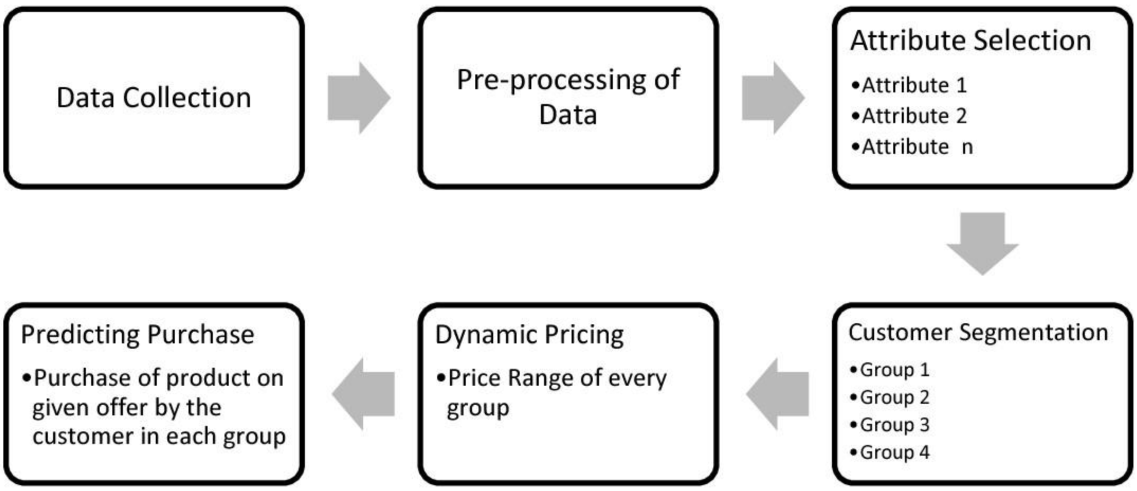
```python
def custom_where(ele):
    '''
    Helper function to combine 3 redundant columns into 1 using simple if-else
    '''

    x,y,z = list(map(str,ele))
    if x == '1':
        if y=='1':
            if z=='1':
                return 'within 25mins'
            else:
                return 'within 15mins'
        else:
            return 'within 5mins'
```

```python
df['driving_distance'] = df[['toCoupon_GEQ5min','toCoupon_GEQ15min','toCoupon_GEQ25min']].apply(custom_where,axis=1,raw=True)
df.drop(['toCoupon_GEQ5min','toCoupon_GEQ15min','toCoupon_GEQ25min'],axis=1,inplace=True)
```

TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that reflects the importance of a term in a document relative to a collection of documents. In the context of coupon purchase prediction, TF-IDF can be used to represent the significance of words or terms within the dataset of customer behavior, purchase history, and coupon usage. By applying TF-IDF, the model can assign weights to terms based on how frequently they appear in a document (term frequency) and how rarely they appear across all documents (inverse document frequency). This can help in identifying important terms and features for predicting coupon redemptions and purchases.



MODELING AND EVALUATING

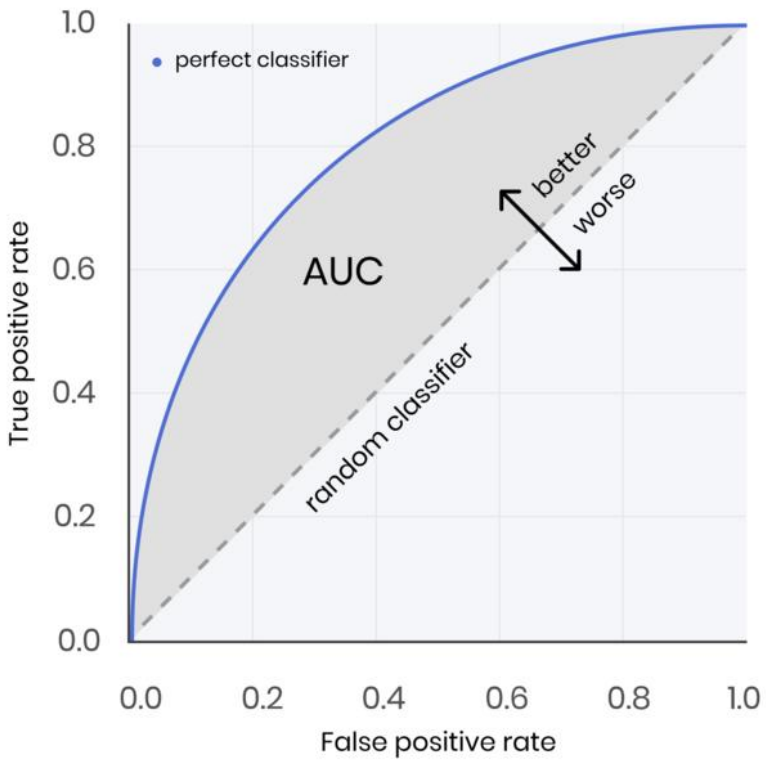
- **MODEL PREDICTION** for Coupon Redemption The random forest model predicted a significant increase in the coupon redemption rate, up to 660%, by increasing X type promotions in the test dataset. This indicates a positive impact on coupon redemption when scaling up certain promotions.

Several machine learning techniques have been employed to tackle this binary classification problem. Tree-based models, such as XGBoost and LightGBM, have proven to be highly effective due to their ability to capture non-linear relationships and handle high-dimensional data 3.
Neural network architectures like Capsule Networks 6 and Multi-armed Bandits combined with clustering techniques (e.g., K-Means) 5 have also been explored for coupon recommendation systems.

TWO-STEP FRAMEWORK for Purchase Prediction A two-step framework involves constructing an undirected graph based on click data to learn a general representation of users and items, leading to a proposed purchase prediction model. This method aims at improving purchase prediction accuracy.
Comprehensive feature engineering plays a vital role in improving model performance. Relevant features can include user demographics, purchase history, coupon usage patterns, product categories, and coupon characteristics (e.g., discount rate, expiration date) 2 4 10.
Advanced techniques like SHAP (SHapley Additive exPlanations) 2 can provide insights into feature importance and model interpretability, thereby enhancing the effectiveness of coupon targeting strategies.

EFFECTIVENESS VERIFICATION of Coupon Usage An effectiveness verification framework analyzes whether a target user will make a purchase using coupons. It includes an architecture diagram for understanding model predictions and evaluating the importance of features to enhance coupon effectiveness.

Typical evaluation metrics for coupon purchase prediction include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC). However, it is essential to consider business objectives and the cost of false positives/negatives when selecting the appropriate metric



BUSINESS CONSTRAINTS

- High Recall and High Precision, both are very important.
- Low latency(prediction) is required as we want to recommend the coupon to the user while he is in his vehicle and in the neighborhood of the venue.
- Interpretability will help but not super important.
- We will tackle this problem by breaking it into 2 subparts i.e. Featurization and Modelling, we'll use a wide range of techniques and conduct lots of experiments to see what works. Let's go step by step.

EXPLORATORY DATA ANALYSIS

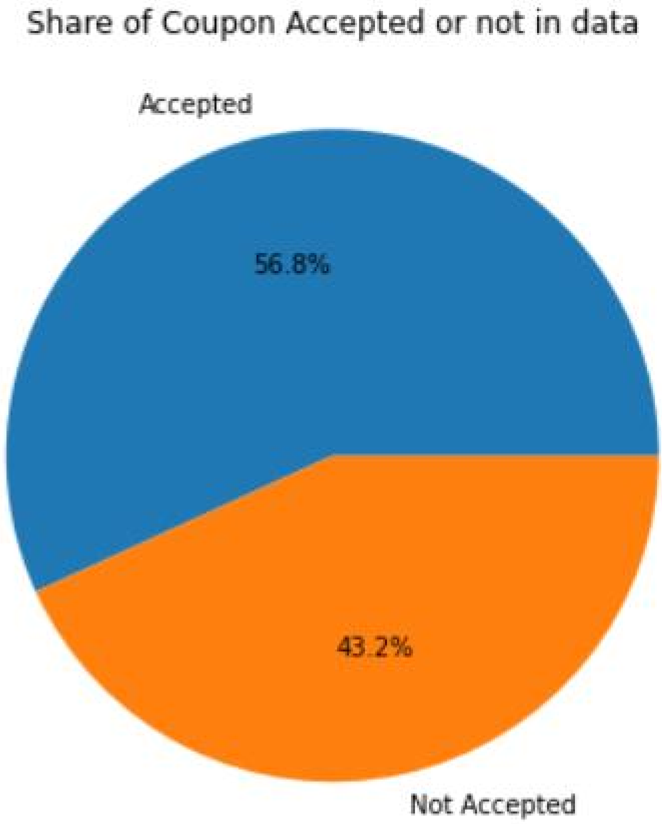
This is the very first but the most important part of our case-study (or any project!). This will give us interesting insights on as to what features are important and mainly what are general trends/patterns which decide whether a person accepts OR rejects a coupon. Let's have a look at the data:

survived	pclass	name	sex	age	sibsp	parch	ticket	fare	cabin	embarked
0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3	7.925		S
1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q

This is a big problem as our data size is too small and we can easily run into problems like Curse of Dimensionality and Overfitting. Moreover, each and every attribute of this dataset is of categorical/discreet nature. That's right, not a single numerical column. Even attributes like Age and temperature had been bucketized before providing the data.

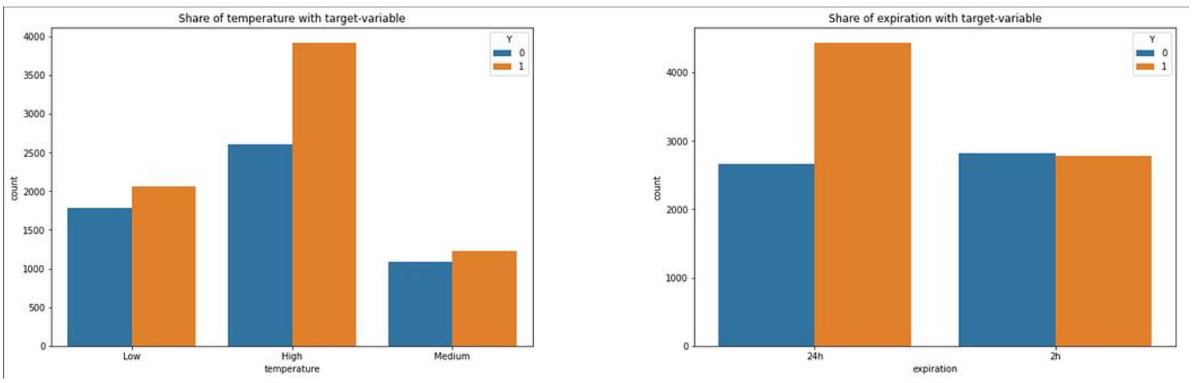
Luckily, we don't have a lot of missing values except of few columns. Car attribute has more than 99% of missing values so we will simply drop it (probably because this must have been an optional question in the survey) as imputing at so many places could risk changing the distribution of data. Other attributes have around 1% missing values only so we'll impute them with mode.

DISTRIBUTION OF CLASSES

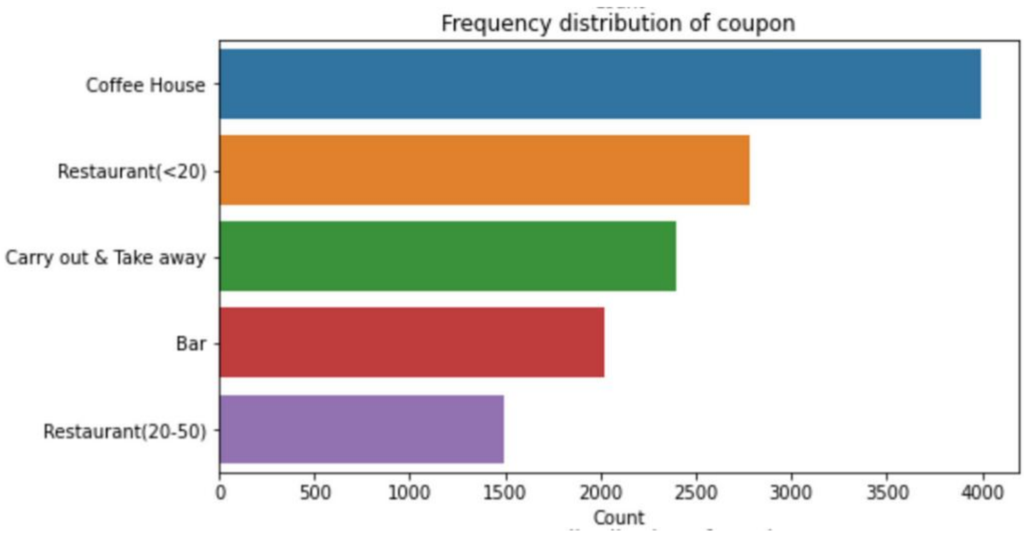


We don’t suffer from severe class imbalance, but balancing can be a part of modelling experimentations.

1. Temperature and Weather effect on target variable

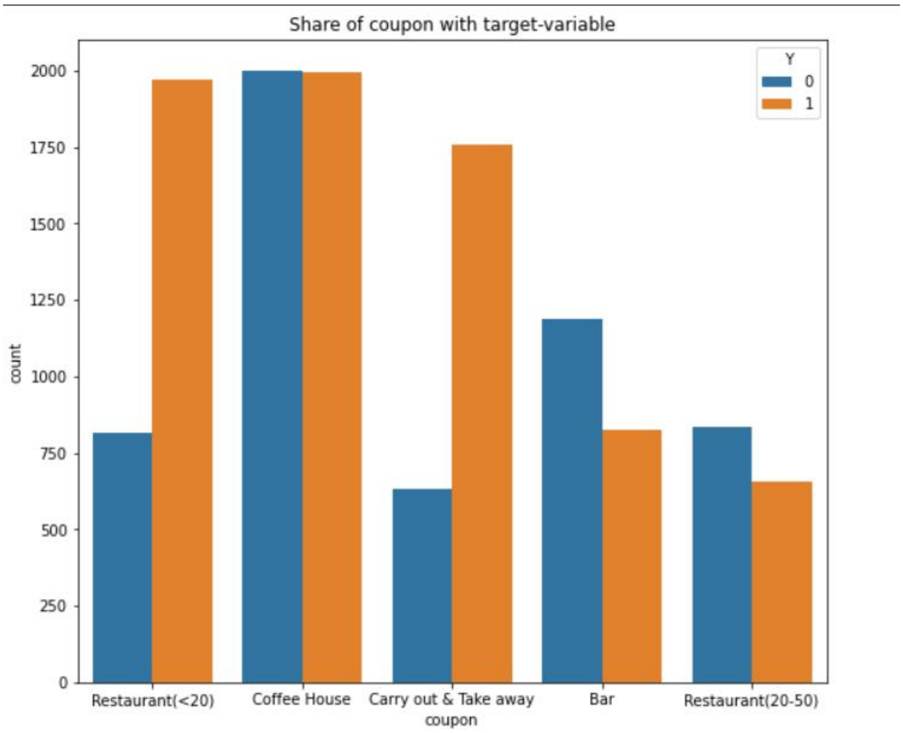


2. Distribution of Coupon types in data



As we can see we have more coupon acceptance when temperature is on the higher side and people have 24hrs before the coupon expires.

3. Coupon type with target



As clearly evident from above plot, Cheap restaurants and carry out/ takeaway have the highest coupon acceptance rates. For coffeehouse we're unsure and for bars and Expensive restaurants, people mostly reject their coupons.

Challenges and Future Directions

One of the key challenges in coupon purchase prediction is handling the dynamic nature of user preferences and market trends. Incorporating temporal dynamics and sequential patterns into the models could improve prediction accuracy 7 15. Furthermore, causal machine learning approaches 4 and counterfactual evaluation techniques could provide a more reliable assessment of the actual impact of coupon campaigns on customer behavior and revenue.

Prediction with Random Forest Model

Changing promotions from type Y to type X in the test dataset increased the coupon redemption rate by 660% according to the random forest model. This suggests that scaling up type X promotions can lead to a significant boost in coupon redemption.

Predicting Coupon Redemption Rates

Predicting coupon redemption rates is a critical task for businesses to optimize their marketing strategies and maximize the effectiveness of coupon campaigns. Several approaches and techniques have been developed to address this challenge.

Machine Learning Models

Machine learning models have proven to be effective in predicting coupon redemption rates. One popular approach is the use of random forest models, which can identify the impact of different promotion types on coupon redemption rates. By analyzing the redemption patterns in the test data, these models can predict the increase in redemption rates by scaling up certain promotions 1.

Two-Step Framework

A two-step framework has been proposed to improve purchase prediction accuracy, which can be applied to coupon redemption prediction. This method involves constructing an undirected graph based on user click data to learn general representations of users and items. These representations are then used to build a purchase prediction model, which can be adapted for coupon redemption predictions 2.

Effectiveness Verification Framework

Another approach is the effectiveness verification framework, which analyzes whether a target user will make a purchase using coupons. This framework includes an architecture diagram to understand model predictions and evaluate the importance of features that contribute to coupon effectiveness. By identifying the most influential factors, businesses can tailor their coupon campaigns to maximize redemption rates 3.

Feature Engineering

Effective feature engineering is crucial for accurate coupon redemption prediction. Features such as customer demographics, purchase history, coupon characteristics (e.g., discount amount, expiration date), and engagement metrics can be used to train machine learning models. Additionally, incorporating external data sources, such as market trends and competitor information, can further enhance the predictive power of the models.

Continuous Monitoring and Optimization

Coupon redemption patterns can change over time due to various factors, such as changing customer preferences, market conditions, and competitor actions. Therefore, it is essential to continuously monitor and optimize the prediction models by incorporating new data and updating the models regularly. This iterative process ensures that the models remain accurate and effective in predicting coupon redemption rates.

CONCLUSION AND FUTURE SCOPE

This study presents a segmentation-based hybrid model for e-coupons' redeemed prediction. Our findings report that there is obvious heterogeneity in consumers' habits about purchasing with e-coupons, including e-coupons types, usage frequency, discount rate, etc., consumers cannot be seen as a homogenous group to distribute e-coupons. In the study, consumers are aggregated into four segments based on their online consumption records, respectively labelled as "potential e-coupons user", "low.

Key Findings in Coupon Purchase Prediction Predictive analytics models can help determine the propensity of customers to redeem coupons by analyzing past online sessions and shopping logs. These models aim to predict the redemption status of a coupon for a specific customer. source

Future Scope in Coupon Purchase Prediction Developing models to improve the coupon redemption rate involves minimizing or maximizing mathematical formulas such as loss functions or information gain formulas. Despite insightful conclusions, there are limitations to this modeling work that need to be addressed in future research. source

Data Insights for E-commerce Coupon Usage Customer purchase behavior in e-commerce can be predicted by analyzing past online sessions and shopping logs. The aim is to predict whether users will purchase specific products during their next visit to the e-commerce platform, emphasizing the importance of data analysis in predicting coupon usage

Predicting coupon purchase behavior is a crucial task for businesses to optimize their marketing strategies and increase revenue. Through the analysis of customer data, purchase history, and coupon redemption patterns, machine learning models can effectively identify customers with a high propensity for coupon usage and tailor promotions accordingly.
The results of these predictive models offer valuable insights into consumer behavior, enabling companies to:
Personalize coupon offers and target the right customers at the right time 1.
Optimize marketing budgets by focusing on customers most likely to redeem

coupons 3.
Enhance customer loyalty and retention through targeted promotions 6.
Understand the impact of various factors (demographic, product category, etc.) on coupon usage 8.
Future Scope
While significant progress has been made in coupon purchase prediction, there are several areas for further exploration and improvement:

Incorporating Real-Time Data: Integrating real-time data sources, such as browsing behavior, location, and social media interactions, can enhance the accuracy of predictive models and enable more dynamic coupon recommendations 5.

Ensemble and Deep Learning Models: Exploring advanced techniques like ensemble modeling and deep learning can capture complex patterns and non-linear relationships, potentially improving prediction performance 11.

Explainable AI: Developing explainable AI models can provide insights into the underlying reasons for coupon purchase decisions, enabling more transparent and trustworthy recommendations 13.

Multi-Objective Optimization: Incorporating multiple objectives, such as revenue maximization, customer satisfaction, and long-term loyalty, can lead to more holistic and sustainable coupon strategies 16.

Privacy and Ethics Considerations: As coupon prediction involves personal data, addressing privacy concerns and adhering to ethical guidelines is crucial for responsible implementation and building consumer trust 17.

By continually advancing coupon purchase prediction models and integrating new data sources and techniques, businesses can stay ahead of the curve and deliver more effective, personalized, and ethical coupon marketing campaigns.

REFERENCES

- https://www.ijsrd.com/articles/IJSRDV7I21146.pdf
- http://efimov-ml.com/pdfs/CouponPurchase2015.pdf
- https://fastercapital.com/content/Coupon-machine-learning--How-to-use-machine-learning-to-analyze-and-predict-coupon-performance-and-behavior.html
- https://www.linkedin.com/advice/3/what-causes-memory-leaks-programming-skills-programming-cqvoe
- https://github.com/pcsanwald/kaggle-titanic/blob/master/train.csv
- http://localhost:8888/edit/Downloads/kaggle-coupon-purchase-prediction-master/kaggle-coupon-purchase-prediction-master/dataset.py
- http://localhost:8888/edit/Downloads/kaggle-coupon-purchase-prediction-master/kaggle-coupon-purchase-prediction-master/make_data.py
- http://localhost:8888/edit/Downloads/kaggle-coupon-purchase-prediction-master/kaggle-coupon-purchase-prediction-master/pairwise_ranking_accuracy.py
- https://github.com/pcsanwald/kaggle-titanic/blob/master/train.csv