Deccan Education Society's
Fergusson College (Autonomous), Pune
Department of Computer Science

A

Project Report
on

Predictive Modeling of Coupon Redemption Behavior Using Machine Learning Techniques

In partial fulfillment of Post Graduate course

in

M.Sc. Data Science – I

(Semester -II)

CSD-560: Data Science Model Building

SUBMITTED BY
Sahil Sahebrao Khedkar (ROLL NO – 246518)
Vedant Anil Gaikwad(ROLL NO - 246527)
Omkar Mukund Kevate(ROLL NO – 246513)

Deccan Education Society's
Fergusson College (Autonomous), Pune
Department of Computer Science

CERTIFICATE

This is to certify that the project entitled Coupon redemption and Prediction submitted by

1. Sahil Sahebrao Khedkar.
2. Vedant Anil Gaikwad
3. Omkar Mukund Kevate

in partial fulfillment of the requirement of the completion of M.Sc. Data Science - I [Semester-II], has been carried out by them under our guidance satisfactorily during the academic year 2024-2025

Place: Pune
Date: / /2025

Project Guide:
Name: Dr. Rasika Kulkarni.

Dr. Kavita Khobragade
Head, Department of Computer Science
Fergusson College (Autonomous), Pune

Examiners Name **Sign**

1. ______________________________ ______________________________

2. ______________________________ ______________________________

Index

Introduction

The competitive scenario with modern retail, discount marketing and coupon -based campaign has proven to be an important tool for customer collection, storage and loyalization of loyalization. Coupons act as a direct encouragement to operate consumer engagement, affect buying decisions and optimize ROI. However, the effectiveness of such campaigns rests on the exact prediction of the customer's redemption behavior. Store suppliers face the challenge with targeted coupon production and design of personal marketing strategies that resonate with different customer segments. The project focuses on developing a forecast model to predict the possibility of coupon redemption, enabled to limit the emergency strategy, allocate resources effectively and maximize the success of the campaign.

Problem Statement:

Retailers use coupon promotions via channels like email and notifications to attract and retain customers. To improve campaign effectiveness, they want to predict whether a customer will redeem a given coupon. Accurate predictions will help design better-targeted marketing strategies. The model's performance will be evaluated using the Area Under the ROC Curve (AUC-ROC) on the test set.

Motivation

In today's competitive retail landscape, businesses frequently offer coupons as a marketing strategy to attract and retain customers. However, not all distributed coupons are redeemed, leading to inefficient resource allocation and missed revenue opportunities. Predicting which customers are most likely to redeem coupons can help businesses target promotions more effectively, reduce marketing costs, and increase customer engagement.

By leveraging data-driven techniques and machine learning, we can uncover patterns in customer behaviour and build intelligent systems that predict coupon redemption likelihood. This enables personalized marketing strategies, better ROI, and improved customer satisfaction. The motivation behind this project is to bridge the gap between marketing efforts and customer behaviour by transforming raw data into actionable insights.

Project Overview

- **Objective**: Predict if customers redeem coupons shared via email, notifications, etc., to refine coupon design and marketing tactics.
- **Data**: Customer demographics, past redemption patterns, coupon attributes (discount %, validity), and campaign channels.
- **Approach**: Machine learning models (e.g., XGBoost) trained to estimate redemption probability, optimized for ROC-AUC scores.
- **Evaluation**: ROC-AUC metric assessed through contest submissions, as test set outcomes are undisclosed.

Workflow

The project follows a systematic data science workflow consisting of the following key stages:

1. Data Collection

- Gather data from available sources, including customer demographics, coupon details, campaign information, and transaction history.

2. Data Preprocessing

- Handle missing values, duplicates, and outliers.
- Encode categorical variables using techniques like Label Encoding or One-Hot Encoding.
- Normalize numerical features to ensure consistency.

3. Exploratory Data Analysis (EDA)

- Perform visual and statistical analysis to understand data distribution, relationships, and trends.
- Identify key features that influence coupon redemption.

4. Feature Engineering

- Create new meaningful features (e.g., average spending, coupon frequency, days since last transaction).
- Select important features using correlation analysis or model-based techniques.

5. Model Building

- Split the dataset into training and testing sets.
- Train a LightGBM (Light Gradient Boosting Machine) model, known for its efficiency and speed with large datasets.
- Leverage LightGBM's ability to handle categorical features and its built-in regularization for improved generalization.

6. Model Evaluation & Tuning

- Compare model performance using cross-validation.
- Optimize hyperparameters using techniques like Grid Search or Random Search.

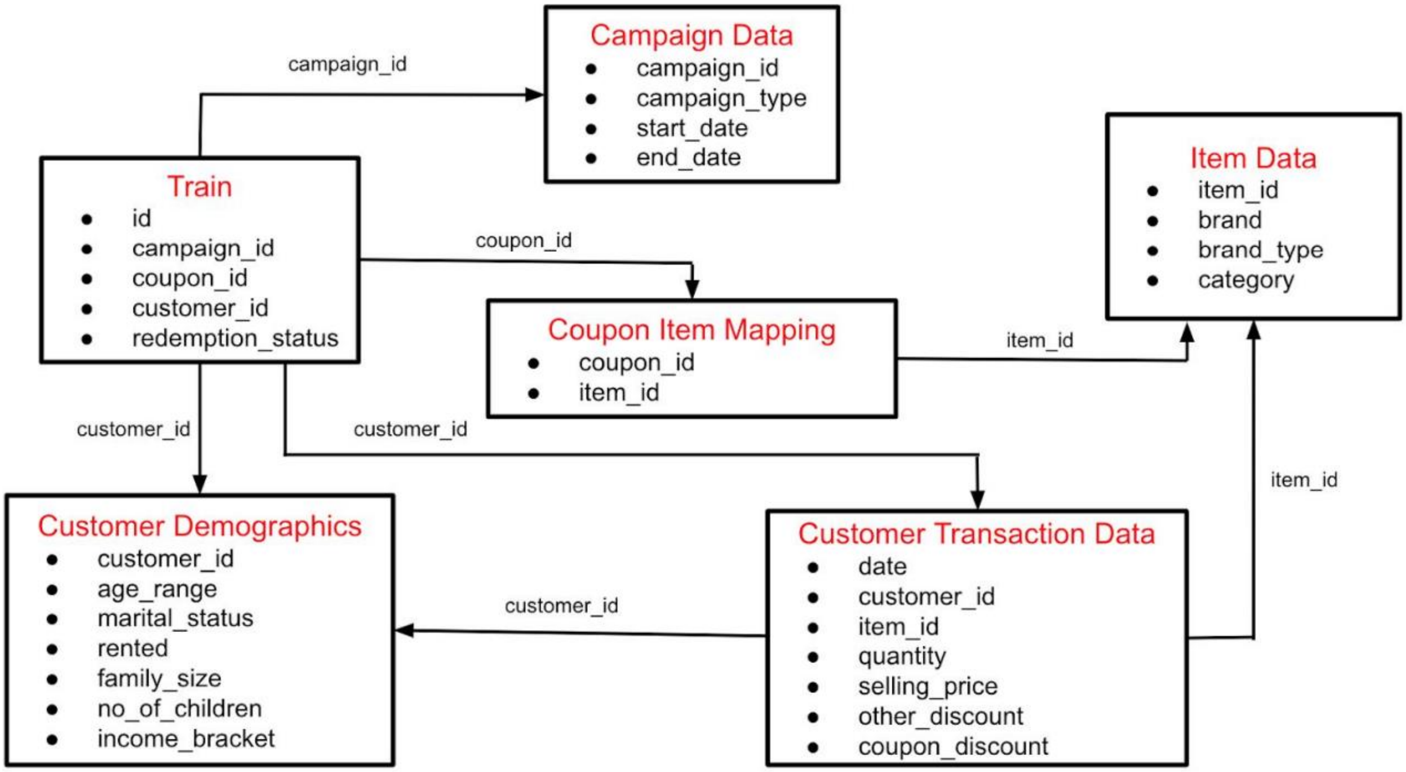
7. Visualization & Dashboard

- Create an interactive dashboard (e.g., using Streamlit) to visualize predictions and key insights.
- Show coupon redemption probabilities, feature importance, and customer segments.

Data source and Data Description

https://www.kaggle.com/datasets/vasudeva009/predicting-coupon-redemption/data
https://www.analyticsvidhya.com/datahack/contest/amexpert-2019-machine-learning-hackathon/

Dataset :



Campaign Data
- campaign_id
- campaign_type
- start_date
- end_date

Train
- id
- campaign_id
- coupon_id
- customer_id
- redemption_status

Item Data
- item_id
- brand
- brand_type
- category

Coupon Item Mapping
- coupon_id
- item_id

Customer Demographics
- customer_id
- age_range
- marital_status
- rented
- family_size
- no_of_children
- income_bracket

Customer Transaction Data
- date
- customer_id
- item_id
- quantity
- selling_price
- other_discount
- coupon_discount

key components:

1. **Customer and Coupon Interaction Data (train.csv, test.csv)**
   - Records of coupons offered to customers during specific campaigns.
   - Includes the target variable (redemption_status) in the training set.
2. **Campaign Details (campaign_data.csv)**
   - Information about different campaigns, including type and duration.
3. **Coupon-Item Mapping (coupon_item_mapping.csv)**
   - Lists which items are valid for discount under each coupon.
4. **Customer Demographics (customer_demographics.csv)**
   - Data on age range, family size, marital status, income, etc.
5. **Transaction Data (customer_transaction_data.csv)**
   - Detailed purchase history of customers, including discounts and prices.

Software and Technology Requirements

1. Programming Language

- **Python**
  Used for data preprocessing, feature engineering, model building, and evaluation due to its flexibility, extensive libraries, and community support.

2. Libraries and Frameworks

- **Pandas, NumPy**
  Utilized for efficient data manipulation, cleaning, and numerical operations.
- **Matplotlib, Seaborn**
  Used for data visualization and exploratory data analysis to uncover trends, patterns, and insights.
- **LightGBM**
  The primary machine learning framework used for training the classification model. It is highly efficient for large datasets and performs well with categorical features.
- **Scikit-learn**
  Employed for model evaluation (e.g., AUC-ROC, precision, recall), train-test data splitting, and supporting preprocessing tasks.

3. Development Environment

- **Visual Studio Code (VS Code)**
  Interactive development environment used for coding, debugging, and project documentation.

4. Version Control & Collaboration Tools

- **Git**
  For tracking code changes, version control, and collaborative development.
- **GitHub**
  Used as the remote repository for storing project files and facilitating team collaboration.

Exploratory Data Analysis (EDA)

Using Matplotlib and Seaborn, several univariate and bivariate analyses were conducted. Bar plots, histograms, box plots helped in identifying:

1) Bar Charts

Used to compare Distribution of Customers, Distribution of Coupons, Distribution of duration, monthly unique customers, monthly Unique Customers, Monthly Unique Items, Monthly Transactions, Monthly transactions Monthly transactions with Coupon Discount following depicts are as follows-

(a) There are 627 (78%) coupons which are not redeemed by any customers

(b) There are 1181 (85%) customers who has not redeemed any coupons

(c) there are not much transaction in the initial 3 months, may be client has just started its business

(d) Month 19 can't be considered in any analysis because it only contains 3 days transactions Customers, Items and total transactions grew in initial 3 months and then remains almost stable Transactions with coupon discount has fluctuations

2) Pie Charts

Used to show the proportion of redeemed vs. non-redeemed coupon

(a) The data is highly Imbalanced

3) Histograms

Show the distribution of numeric variables such as customer income, total purchases, and coupon discount value.

(a) Helps detect skewness and identify outliers.

4) Box Plots

Used to visualize the spread and outliers in features like transaction amount.

(a) Useful for understanding variability within customer behaviour.

5) ROC Curve

Used to evaluate model performance by plotting the True Positive Rate vs. False Positive Rate.

(a) AUC score under the ROC curve indicates model accuracy.

b. Technologies Used for Visualization

Python (Matplotlib & Seaborn) – For EDA charts like bar graphs, histograms, and box plots.

Plotly – For interactive visualizations, especially helpful in dashboard integration.

Scikit-learn & Matplotlib – For plotting the ROC-AUC curve

Data Preprocessing

Data Cleaning:

After analysing the dataset, below issues were found. Missing Data For some customers, "no_of_children" and "marital_status" fields are missing. For “no_of_children”, making an assumption that if it is NaN then it means that customers have no/zero children. Hence, filling missing “no_of_children” with zero. Marital status of the customers was calculated using family size and no of children. If a customer has a member who is not his/her child than consider that extra member to be his/her spouse. Hence, If there is an extra member in the customers family then that customer is married. Train and Test set contains many customers whose information is not available. Keeping variables to be NaN for the prediction algorithm to handle them. Outliers Data contains many outliers in the customer’s transactions. Since, there are many outliers, Prediction algorithm needs to be trained with outliers.

Data Merging:

Coupon Information

Since, merging coupon information with training and test data is many-to-many mapping, it needs to be reduced to one-to-many mapping. Extracting below summary variables from coupon information.
c_unique_items, c_unique_brand, c_freq_brand, c_rare_brand, c_items_freq_brand, c_items_rare_brand, c_unique_brandt, c_freq_brandt, c_rare_brandt, c_items_freq_brand, c_items_rare_brand, c_unique_category, c_freq_category, c_rare_category, c_items_freq_category etc

Customer transactions

Similar to coupon information, extracting below summary variables from customer transactions considering them as the customers buying behaviour.
Overall_unique_items, overall_items, overall_quantity, overall_sprice, overall_bprice, overall_odiscount, overall_cdiscount, overall_tdiscount, overall_sprice_pq, overall_bprice_pq, overall_odiscount_pq, overall_cdiscount_pq, overall_tdiscount_pq, overall_unique_brand, overall_freq_brand, overall_rare_brand,

Campaign and Customer Information

Campaign information and customer information both have one-to-many mapping with train and test data. Hence, they can be merged easily using left join.

Feature Engineering:

- Summary statistics like frequency, brand coverage, discount patterns were created.

Deriving Features :

Deriving more features which represents the change in customer behaviour:
diff_range_unique_items, diff_range_items, diff_range_quantity, diff_range_sprice, diff_range_bprice, diff_range_odiscount, diff_range_cdiscount, diff_range_tdiscount, 8 diff_range_podiscount etc

Model Building

The project employed LightGBM (Light Gradient Boosting Machine), a high-performance gradient boosting framework particularly suitable for:

- Large-scale datasets,
- Handling categorical features directly,
- Efficient training with lower memory usage.

The modelling focused on capturing complex patterns in the data by using a feature-rich approach, combining:

- Customer demographics and behaviour
- Coupon attributes,
- Campaign metadata.

To further boost accuracy, hyperparameter tuning and ensemble learning techniques were applied.

Model Outcomes

The LightGBM model successfully predicted coupon redemption as a binary classification task (0 or 1), where:

- 1 = Coupon Redeemed
- 0 = Coupon Not Redeemed

Key model outcomes included:

- High predictive performance with strong generalization.
- Robust handling of missing values and noisy features.
- Insight into feature importance, highlighting customer behaviour as a critical driver of redemption.

Evaluation:

The primary evaluation metric was AUC-ROC, due to its effectiveness in binary classification and imbalanced data scenarios.

Model Configuration	**Validation AUC**	**Test AUC**
Base LightGBM Model	0.9970	0.9016
After Hyperparameter Tuning	0.9970	0.9060
Final Ensemble Model	—	0.9132
Private Leaderboard (Benchmark)	—	0.8979

Comparison with Existing Systems / Models

Comparison Scenario	Test AUC
Without customer behaviour features	0.8690
With customer behaviour features	0.9016

Models without customer behaviour features saw a significant drop in performance, demonstrating their importance.

The ensemble approach, averaging LightGBM models with different num_leaves (15, 20, 25), provided the best test performance (AUC = 0.9132), proving its effectiveness over individual models.

Deployment

User Interface for Deployment (Streamlit)

The user interface was developed using Streamlit, an open-source Python library that allows for the rapid creation of interactive web applications for data science and machine learning workflows.

Key Components of the Streamlit App

1. Data Upload

- Users can upload a CSV file containing new customer or campaign data.
- The app reads and previews the first few rows using st.file_uploader() and st.dataframe().
- Basic validation is applied to ensure required columns are present.

2. Prediction Module

- Upon upload, the data is pre-processed using the same pipeline as the training phase.
- The trained LightGBM model is loaded using joblib or pickle.
- Predictions (redeemed/not redeemed) are displayed in tabular format with:

3. Visual Analytics

- Integrated Matplotlib and Seaborn visualizations rendered using st.pyplot():
  - ○ Coupon redemption distribution
  - ○ Feature importance from the model
  - ○ Customer segment-wise analysis (e.g., by age, income)

4. Download Predictions

- Users can download the prediction results as a CSV file using st.download_button(), enabling further offline analysis or reporting.

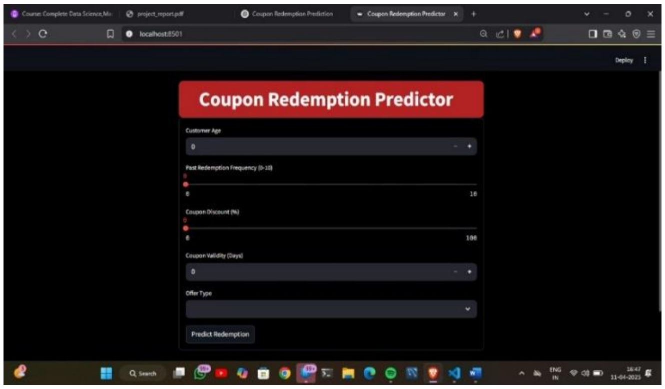
6. Deployment

- The Streamlit app can be hosted on:

  Streamlit Cloud

Input screen

output screen


Course: Complete Data Science,Ma
project_report.pdf
Coupon Redemption Prediction
Coupon Redemption Predictor
localhost:8501
Deploy
Search
ENG
IN
16:47
11-04-2025

Coupon Redemption Predictor

Customer Age: 0

Past Redemption Frequency (0-10): 0 (0 – 10)

Coupon Discount (%): 0 (0 – 100)

Coupon Validity (Days): 0

Offer Type:

Predict Redemption

Drawbacks and Limitations

- **Missing Demographic Data**

  Nearly 50% of customers had incomplete demographic information, which limits the model's ability to personalize predictions.

- **Imbalanced Dataset**

  The dataset is highly skewed, with a large majority of coupons not being redeemed. This imbalance can affect model performance and requires careful handling.

- **Static Feature Window**

  Customer behaviour was analysed within a fixed time window (223 to 110 days before a campaign). A dynamic or campaign-specific window might capture behaviour changes more accurately.

- **Limited External Features**

  The dataset does not include contextual variables like time of day, location, or competitor activity, which could influence redemption behaviour.

- **Assumption-Based Imputation**

  Some missing values were filled using assumptions (e.g., no children = 0, marital status inferred from family size), which might introduce bias.

- **No Confusion Matrix Analysis**

  Evaluation focused mainly on AUC; other important metrics like precision, recall, and confusion matrix were not analysed, which could help assess model bias.

- **Model Interpretability**

  Ensemble models like LightGBM, while powerful, are less interpretable. This can make it harder for marketing teams to understand why a customer is likely to redeem a coupon.

- **Overfitting Risk**

  Very high validation scores compared to test scores suggest possible overfitting during training and hyperparameter tuning.

Future Enhancements

To further improve the model's accuracy, usability, and real-world effectiveness, the following future enhancements are proposed:

Enhanced Missing Data Handling

  - Apply advanced imputation methods such as k-NN, matrix factorization, or deep learning-based imputers to predict missing demographic or behavioural data more accurately.

- **Dynamic Time Windowing**
  - Implement personalized or campaign-specific time windows for behavioural feature extraction, improving the relevance and responsiveness of features to customer actions.
- **Threshold Tuning & Error Analysis**
  - Go beyond ROC-AUC by fine-tuning decision thresholds using Precision, Recall, and F1-score to minimize false positives/negatives based on business impact.
  - Visualize model decisions using Confusion Matrix for better diagnostic insights.
- **Model Interpretability**
  - Integrate explainability tools like SHAP (Shapley Additive explanations) or LIME to quantify feature impact and build trust among business stakeholders.
- **Enrichment with External Data**
  - Incorporate contextual signals such as store location, regional holidays, weather data, or competitor promotions to enhance model depth and real-world adaptability.
- **Advanced Ensemble Techniques**
  - Explore model stacking, blending, and hybrid ensembles that combine LightGBM, XGBoost, CatBoost, and potentially neural networks to leverage strengths of multiple algorithms

.

Conclusion

This project successfully developed a high-performing predictive model for coupon redemption classification using a combination of customer behaviour, coupon metadata, and campaign details. The use of LightGBM, along with thoughtful feature engineering and hyperparameter tuning, resulted in a strong test AUC of 0.9132, outperforming baseline approaches.

The model demonstrates significant potential in helping retailers:

- Target the right customers with personalized offers,
- Optimize marketing campaigns, and
- Reduce promotional waste.

While the model is robust, there is room for future enhancement—particularly in data enrichment, explainability, and error handling. With these improvements, the system could evolve into a powerful real-time marketing assistant, delivering actionable insights for customer engagement and revenue growth.

Bibliography

- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017)
  LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.
- Little, R. J. A., & Rubin, D. B. (2002).
  Statistical Analysis with Missing Data (2nd ed.). Wiley.
- Neslin, S. A., & Clarke, D. G. (1988)
  Sales Promotion: State of the Art. Journal of Marketing, 52(4), 61–71.
- Guyon, I., & Elisseeff, A. (2003)
  . An introduction to variable and feature selection. Journal of Machine Learning Research, 3(Mar), 1157–1182.
- Dietterich, T. G. (2000).
  Ensemble methods in machine learning. In Multiple Classifier Systems (pp. 1–15). Springer.
- Gensch, D. H., & Wellman, B. J. (1997).
  Markets, segments, and all that: Segmentation, target marketing, and positioning. Handbook of Marketing.