In-Vehicle Coupon Recommendation System

Achala Shenoy, Ankita Shukla, Navya Pragathi Molugaram, Sanidhya Mathur

I. Abstract

This project entails an in-depth exploration of a consumer behavior dataset, employing extensive Exploratory Data Analysis (EDA) and implementing various machine learning algorithms. The dataset encompasses 26 columns, including demographic information, preferences, and responses to promotional coupons. Through meticulous data cleaning, preprocessing, and feature engineering, the project aims to uncover patterns, optimize feature selection, and develop predictive models.

II. Introduction

1. Problem Statement

In the realm of marketing strategy, understanding consumer behavior is paramount. The issuance of coupons as an incentive for patronage is a common tactic, but its effectiveness varies widely. Recognizing the intricate interplay of factors that influence coupon redemption can significantly enhance the efficiency of such campaigns. This project addresses the challenge of unraveling the dynamics at play when consumers decide to redeem coupons.

2. Project Goal

The primary goal of this project is to develop a **predictive model** capable of discerning the likelihood of coupon redemption based on a myriad of contextual factors. By comprehensively exploring and **preprocessing** the dataset, applying robust **machine learning** algorithms, and **fine-tuning** model parameters, the intention is to create a tool that can assist marketers in **optimizing** coupon distribution **strategies**. The project seeks not only to predict consumer responses accurately but also to offer insights into the key features influencing these responses.

3. Scope

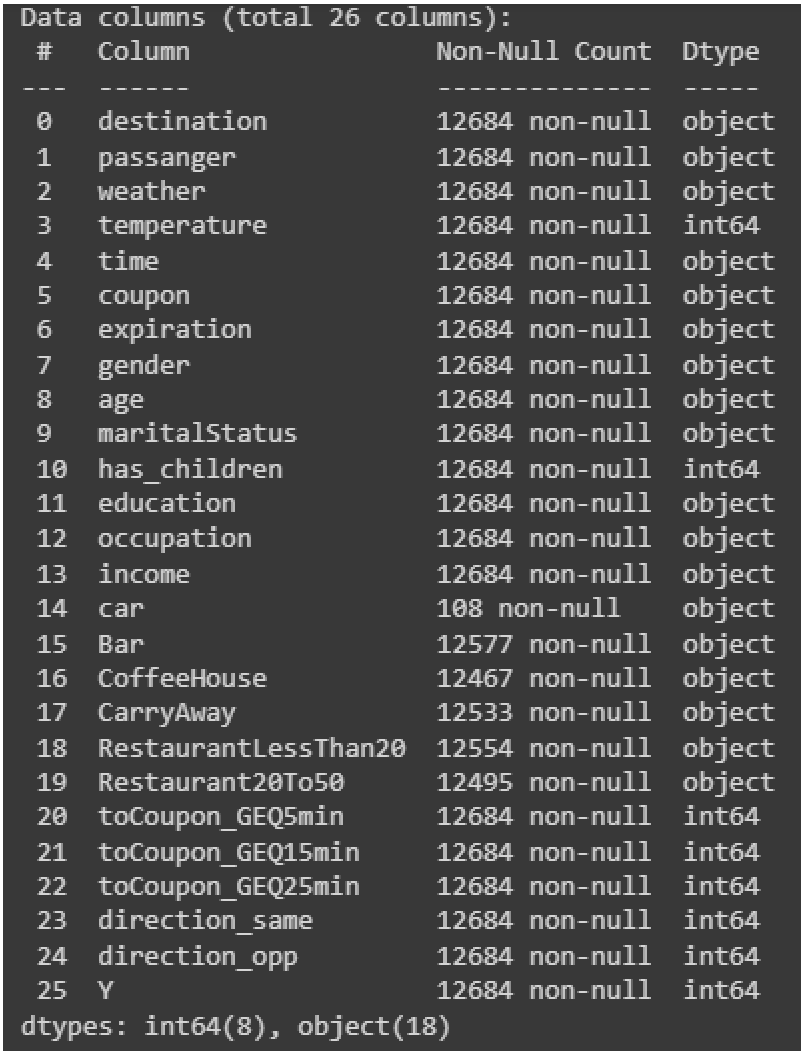
Feature engineering plays a pivotal role in enhancing the interpretability of the model, allowing for actionable insights. Furthermore, this project evaluates the performance of various machine learning algorithms, delving into their strengths and weaknesses in the context of coupon redemption prediction.

4. Significance

Understanding the factors that drive consumer behavior concerning coupon redemption has far-reaching implications for businesses in the dining industry. It enables targeted marketing efforts, resource optimization, and the crafting of more personalized and effective promotional campaigns. The insights derived from this project have the potential to contribute to the refinement of marketing strategies, resulting in increased customer engagement and satisfaction.

III. Data Description

The dataset comprises 12,684 entries with 26 columns. The columns represent various aspects of consumer behavior related to coupon redemption.
Below is a summary of key information:



```
Data columns (total 26 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   destination           12684 non-null  object
 1   passanger             12684 non-null  object
 2   weather               12684 non-null  object
 3   temperature           12684 non-null  int64
 4   time                  12684 non-null  object
 5   coupon                12684 non-null  object
 6   expiration            12684 non-null  object
 7   gender                12684 non-null  object
 8   age                   12684 non-null  object
 9   maritalStatus         12684 non-null  object
 10  has_children          12684 non-null  int64
 11  education             12684 non-null  object
 12  occupation            12684 non-null  object
 13  income                12684 non-null  object
 14  car                   108 non-null    object
 15  Bar                   12577 non-null  object
 16  CoffeeHouse           12467 non-null  object
 17  CarryAway             12533 non-null  object
 18  RestaurantLessThan20  12554 non-null  object
 19  Restaurant20To50      12495 non-null  object
 20  toCoupon_GEQ5min      12684 non-null  int64
 21  toCoupon_GEQ15min     12684 non-null  int64
 22  toCoupon_GEQ25min     12684 non-null  int64
 23  direction_same        12684 non-null  int64
 24  direction_opp         12684 non-null  int64
 25  Y                     12684 non-null  int64
dtypes: int64(8), object(18)
```

1. Data Types

The dataset includes 8 numerical columns (`int64`) and 18 categorical columns (`object`) including the target variable. Examples of categorical columns include 'destination', 'passenger', 'weather', 'time', 'coupon', and 'gender', among others.

2. Summary Statistics for Numerical Columns

- **'temperature'** has a mean of 63.30, ranging from 30 to 80.
- **'has_children'** is a binary variable with a mean of 0.41.
- **'toCoupon_GEQ5min'** has a constant value of 1 and can be considered for removal.
- **'direction_same'** and **'direction_opp'** are binary variables indicating whether the coupon destination is in the same or opposite direction of the driver's position.

3. Missing Values

- Columns with missing values may require imputation or further investigation.

Sr. No	Column Name	Number of Missing Values
1	Car	12576
2	Bar	107
3	CoffeHouse	217
4	CarryAway	151
5	RestaurantLessThan20	130
6	Restaurant20To50	189

4. Unique Values in Each Column

- Categorical columns like 'destination', 'passenger', 'weather', 'coupon', 'maritalStatus', 'education', 'occupation', 'income', and others exhibit diverse categories.
- Some columns, like 'toCoupon_GEQ15min' and 'toCoupon_GEQ25min', are binary.
- Target attribute Y - 1 accepted; Y - 0 Rejected

This overview provides a foundation for exploratory data analysis (EDA) and subsequent model development. Addressing missing values and considering feature engineering could enhance the dataset for effective modeling.

IV. Methods

We focussed primarily on building and evaluating predictive models for coupon redemption. The methodology involves several key steps, including data cleaning, feature engineering, and model selection. Below is an overview of the methods employed in this project:

1. Data Cleaning

- **Handling Missing Values**: The dataset was examined for missing values, and strategies were implemented to address them. Columns with a significant number of missing values (like “Car”) were dropped and other columns with missing values less than 1% were filled using mode imputation.
- **Column Removal**: Columns deemed irrelevant or redundant, such as 'toCoupon_GEQ5min' were removed from the dataset to streamline the feature set.

2. Feature Engineering

- **Categorical Variable Encoding**: Categorical variables were encoded to numerical format using techniques like label encoding and one-hot encoding. This step is crucial for machine learning algorithms that require numerical input.
- **Ordinal Variable Encoding**: Ordinal variables were appropriately encoded to capture the ordinal relationships between categories.
- **Age and Occupation Grouping**: The 'age' and 'occupation' columns were grouped to create more manageable and interpretable categories, enhancing the model's ability to capture patterns.

3. Exploratory Data Analysis (EDA)

- **Statistical Analysis**: Descriptive statistics and data visualization techniques were employed to gain insights into the distribution of key variables, identifying potential patterns or anomalies.
- **Correlation Analysis**: The correlation matrix was examined to understand relationships between variables and identify potential multicollinearity.

4. Model Selection

- **Logistic Regression**: A logistic regression model was employed, given its interpretability and suitability for binary classification problems.
- **Naive Bayes**: The Naive Bayes classifier was chosen for its simplicity and effectiveness, especially in scenarios with categorical features.
- **Neural Network**: A neural network model was implemented to capture complex, non-linear relationships within the data.
- **K-Nearest Neighbors (KNN)**: KNN was utilized to consider the local context of data points in predicting coupon redemption.

5. Evaluation Metrics

- Model performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. These metrics provide a holistic view of the model's effectiveness in predicting coupon redemption.

The model selection process involved training and evaluating each model on the dataset, considering factors such as interpretability, computational efficiency, and predictive performance. This comprehensive methodology ensures a robust exploration of various modeling techniques to identify the most suitable approach for the given problem.

V. Exploratory Data Analysis (EDA)

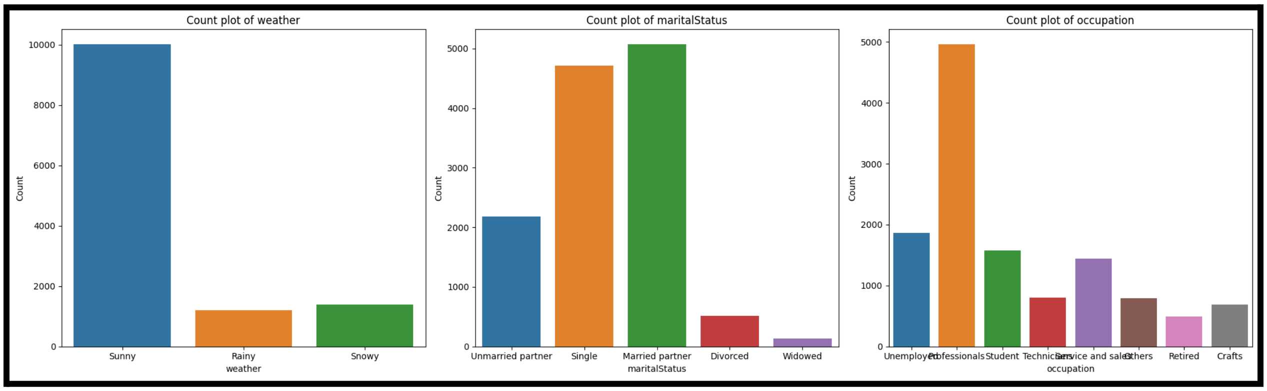
EDA helped understand the dynamics of consumer decisions regarding coupon redemption. It formed the basis for subsequent analyses and modeling. The primary goals of EDA are to unravel patterns, relationships, and anomalies within the data, enabling informed decision-making and hypothesis generation.

1. Distribution of classes

The dataset is partially balanced with acceptance class labels as 57% and rejection class labels as 43%.

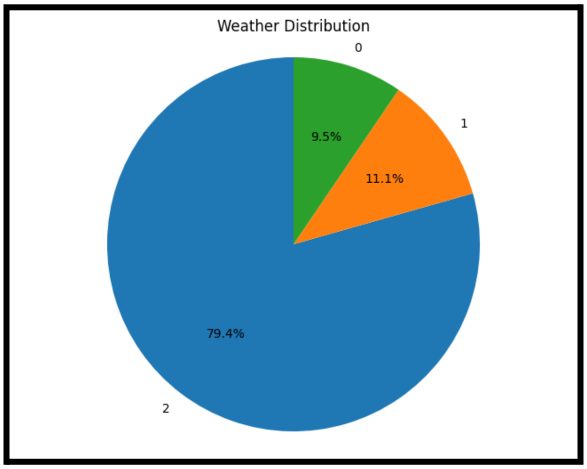
2. Count Plots for Categorical Columns

Count plots were utilized to visualize the distribution of categorical variables such as weather, marital status, and occupation. This technique provides a quick overview of the frequency of different categories, offering insights into the dominant factors within the dataset.



Marital Status: Unmarried partners, singles and married partners form majority of the dataset As compared to the relatively lesser number of divorced and widowed individuals.

Occupation: Professionals, Service and sales workers, students, technicians Retired individuals, and those in crafts exhibit diverse occupational backgrounds

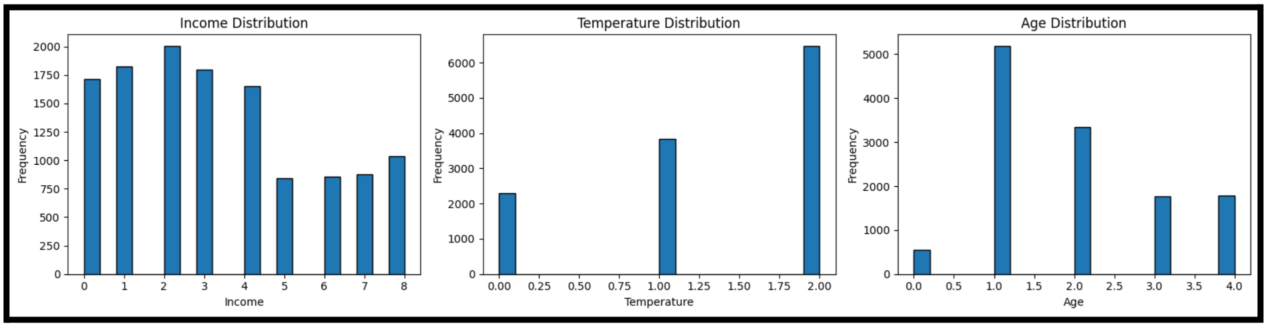


.

Weather: The majority of instances (79.4%) feature sunny weather, followed by rainy (11.1%) and snowy (9.5%) conditions.

3. Histograms for Numerical Columns

Histograms were employed to showcase the distribution of numerical features like income, temperature, and age. These visualizations aid in identifying patterns and central tendencies within each numerical variable, crucial for understanding the diversity and range of the data.



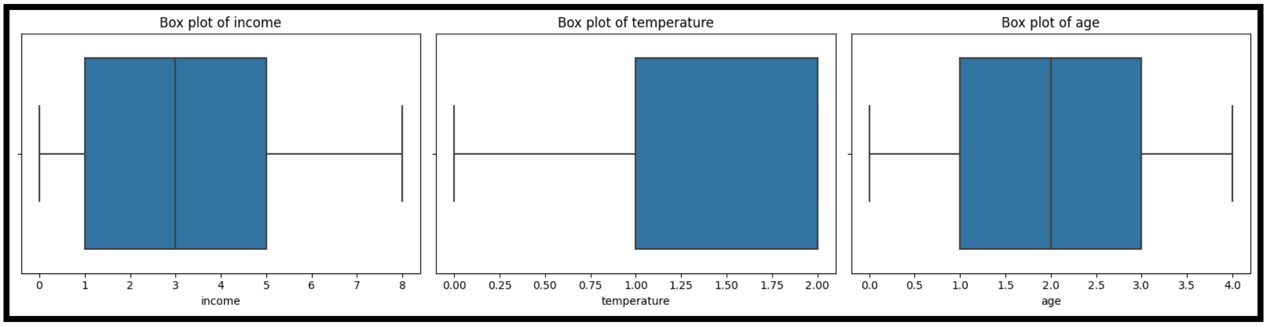
Income Distribution: Peaks are observed at income levels 2, 3, and 4, with variations across the income spectrum.

Temperature Distribution: Three temperature categories (0, 1, 2) showcase frequency distributions, with temperature 2 being the most prevalent.

Age Distribution: Spread across age categories 1 to 4, with age 1 scoring highest frequency.

4. Box Plots for Numerical Columns

Box plots were used to depict the spread and central tendency of numerical features. This technique provides a visual summary of the distribution, highlighting outliers and the variability of the data.



5. How EDA Helped in this Dataset

Pattern Identification:

Count plots for categorical variables revealed patterns in weather conditions, marital status, and occupation. This aids in understanding the prevalent **scenarios under which coupons are redeemed**.

Understanding Income Dynamics

Histograms and box plots facilitated a nuanced understanding of **income distribution**, identifying income levels with higher coupon redemption frequencies. This information is valuable for **targeted marketing strategies**.

Temperature, Weather and Age Insights

Histograms provided insights into the distribution of temperature and age categories, assisting in identifying weather and temperature conditions and age groups more inclined towards coupon redemption.

In summary, the EDA techniques provided a comprehensive understanding of the dataset's composition, unveiled patterns, and guided subsequent steps in the data science pipeline. This foundational analysis contributes to the formulation of hypotheses, feature engineering strategies, and the eventual development of accurate predictive models.

VI. Feature Engineering, Feature Selection and PCA

Feature engineering tailors the features to enhance model performance, feature selection focuses on relevant attributes, statistical analysis provides insights into data patterns, and PCA reduces the dimension of data. Their combined application aims to improve the overall efficiency and effectiveness of the machine learning process

1. Feature engineering

1. **Label Encoding for Categorical Variables**: The categorical variables in the dataset were encoded into numerical values using label encoding. This transformation ensures that non-numeric data is compatible with machine learning models.

2. **Summarization of Age and Occupation Columns**: The 'age' and 'occupation' columns were summarized into fewer classifications, contributing to a more streamlined and meaningful representation of these features.

2. Feature Selection and Statistical Analysis:

To enhance model performance by selecting the most relevant features - PCA was employed.

Principal Component Analysis (PCA):

1. **Centering the Data**: Subtracting the mean, making the data zero-centered.

2. **Covariance Matrix**:Representing the relationships between different features.

3. **Eigen decomposition**: To obtain eigenvalues and eigenvectors.

4. **Sorting Eigenvalues and Eigenvectors**:Descending order to identify the principal components.

5. **Selecting Top Eigenvectors**: A specified number of top eigenvectors corresponding to the largest eigenvalues were selected. In this case, 14 components were chosen.

6. **Transformation**: The original data was transformed using the selected eigenvectors, resulting in a reduced set of features.

7. **Explained Variance Ratio**: The explained variance ratio provides insights into the proportion of the dataset's variance captured by each principal component.

8. **Sum of Explained Variance Ratio**: The sum of the explained variance ratio indicates the cumulative proportion of variance retained in the dataset after dimensionality reduction. In this case, **approximately 93.7% of the variance was retained**.

3. How PCA Helped in this Dataset

1. **Dimensionality Reduction**: PCA facilitated the reduction of the dataset's dimensionality, representing the original features in a more compact form while retaining essential information.

2. **Noise Reduction**: By focusing on the principal components with the highest variance, PCA helps mitigate the impact of noise and less informative features.

3. **Visualization**: Reduced-dimensional data is more amenable to visualization, enabling a better understanding of patterns and relationships within the dataset.

4. **Efficient Modeling**: A dataset with reduced dimensionality can lead to more efficient model training, especially when dealing with high-dimensional data.

VII. Model Explanation

a. Naïve Bayes

The Naïve Bayes classifier operates on the principle of **Bayes' theorem** to perform probabilistic classification. In the context of our machine learning model, we denote the input data as B and the class as A. The classifier evaluates the **probability of the input data belonging to a specific class** by considering the observed values of a set of features or parameters (represented as B in Bayes' theorem).

Results

Accuracy	60.3%
Recall	62.5%
Precision	78.0%
F1 Score	69.4%

Model Selection

- Naive Bayes is chosen for its simplicity, efficiency, and ability to handle categorical data.
- It's particularly suitable for datasets with a large number of categorical features.

b. Logistic Regression

Logistic regression is a statistical method used for predicting the **probability of a binary outcome**, which means an event that has two possible results, like **true** or **false**, **0** or **1**. It's a valuable tool in **classification problems**, where the goal is to assign new data points to specific categories. It essentially helps us make informed decisions about the likelihood of different outcomes of the data

Results

Accuracy	59.5%
Recall	65.6%
Precision	59.9%
F1 Score	62.1%

Model Selection

- Logistic Regression is chosen for its simplicity, interpretability, and effectiveness in binary classification tasks.
- It's particularly suitable for problems where the relationship between features and the target variable is approximately linear.

Model Comparison with Base Model (Naive Bayes)

- Logistic Regression shows a **slightly higher accuracy** compared to Naive Bayes (59.5% vs. 60.3%).
- Precision, Recall, and F1 Score are comparable between the two models, with Logistic Regression having a **marginal advantage**.

c. Neural Network

A neural network is used for pattern recognition, classification, regression, and decision-making. The neural network consists of interconnected nodes which are organized into layers – an input layer, one or more hidden layers, and an output layer. We have applied it as a predictive model to discern complex, non-linear relationships within the dataset.
This involves the **training** of the network on input data, **learning patterns** and **associations**, and subsequently making predictions or classifications based on its acquired knowledge.

Results

Accuracy	57.8%
Recall	57.8%
Precision	100%
F1 Score	73.2%

Model Selection

- Neural networks are chosen to capture complex, non-linear relationships in the data.
- The architecture is kept simple due to the relatively small dataset size.

Model Comparison with Base Model (Naive Bayes)

Neural Network performs slightly worse than Naive Bayes in terms of accuracy..

- The **model excels in recall**, capturing all positive instances, but at the cost of precision, indicating potential **overfitting to the training data**.

d. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a **non-parametric** machine learning algorithm, thus, it does not make any assumptions about the **distribution of the data**. This makes KNN a very flexible algorithm, but it also means that it is more prone to overfitting.

Additionally, KNN is a type of **instance-based learning** where the model makes predictions based on the similarity of input data points. In the context of our project, KNN is applied to **predict coupon redemption** by considering the features and characteristics of **consumers**.

Advantages of KNN:

- Simplicity and ease of implementation.
- Adaptability to different types of data and feature spaces.
- Ability to capture non-linear relationships.

Considerations:

- Sensitivity to the choice of distance metric and K value.
- Computationally intensive for large datasets.

Results

Accuracy	64.9%
Recall	64.9%
Precision	64.8%
F1 Score	64.9%

Model Selection

KNN is chosen for its simplicity and ability to capture local patterns. It's suitable for datasets where instances of the same class tend to cluster together.

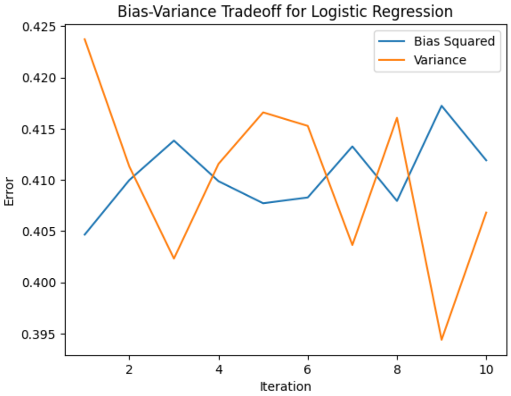
Model Comparison with Base Model (Naive Bayes)

- KNN outperforms Naive Bayes in terms of **accuracy** and **F1 Score**.
- It provides a **more balanced performance** across precision and recall compared to Naive Bayes.

VIII. Bias-Variance Tradeoff

The bias-variance tradeoff helps to balance two types of errors – bias and variance – to achieve optimal model performance. Indicates finding the **right level of model complexity** to achieve a balance between bias and variance. A model that is **too simple (high bias)** may not capture the complexities of the data, while a model that is **too complex (high variance)** may fit the training data too closely and fail to generalize to new, unseen data.

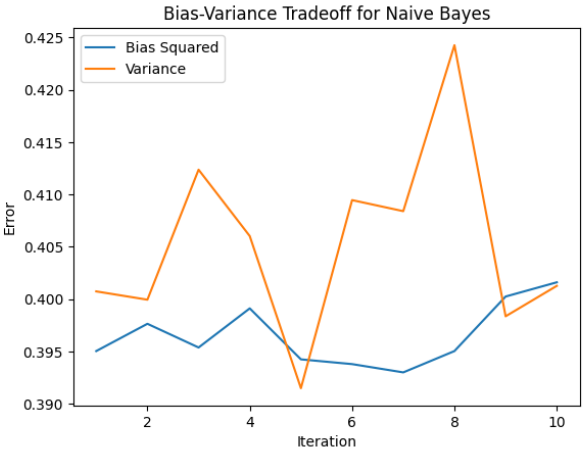
1. Logistic Regression



The **training error** of the logistic regression model **decreases as the number of iterations increases**. This is because the model is learning more and more from the training data.

However, the **test error** also starts to **decrease** at first, but then it starts to **increase** again after a certain number of iterations. This is because the model is starting to overfit the training data and is no longer able to generalize to new data.

2. Naive Bayes

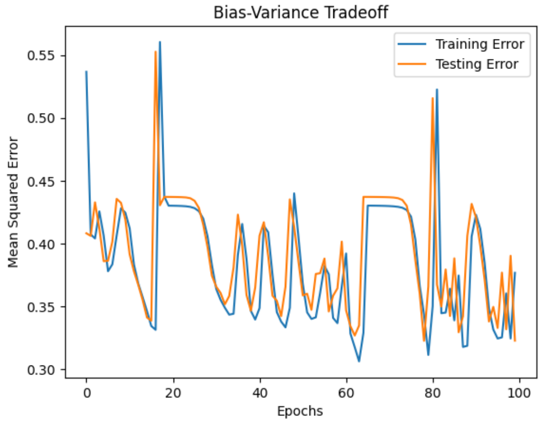


The **training error** and **testing error** of the Naive Bayes Model both decrease as the number of features increases. This is because **increasing the number of features** makes the model more complex and able to fit the training data better.

However, it also **increases the variance of the model**, which means that it is more likely to overfit the training data and perform poorly on new data.

As the number of iterations increases, the **bias increases** as the model assumes independence between the features. The fluctuations in the variance in the acceptable range shows that a model maintains a good balance between stability and flexibility.

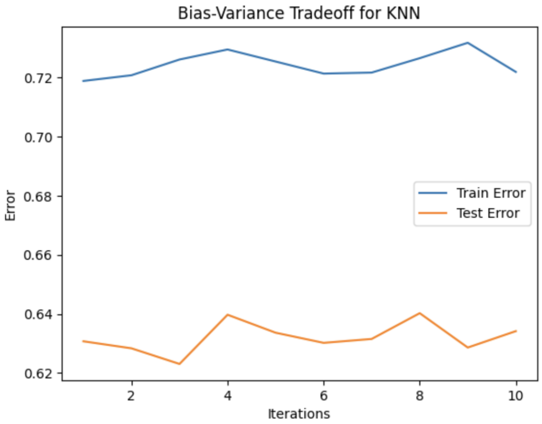
3. Neural Network



The neural network's bias-variance tradeoff graph illustrates **mean squared error values** during training and testing over different **epochs**. The **training error** of the neural network model **decreases** as the **number of epochs increases**. This is because the model is learning more and more from the training data.

However, the **test error** also starts to decrease at first, but then it starts to increase again after a certain number of epochs. This is because the model is starting to overfit the training data and is no longer able to generalize to new data. A **close alignment** between training and testing errors indicates a **well-balanced model**.

4. KNN



The **train error** and **test error** of the KNN model both **decrease** as the number of neighbors (k) increases. This is because **increasing k** makes the **model more complex** and able to fit the training data better.

However, it also increases the variance of the model, which means that it is more likely to overfit the training data and perform poorly on new data.

In general, bias-variance tradeoff graphs help in diagnosing potential issues with the models. A well-tuned model like Neural Networks strike a balance between bias and variance, leading to good generalization on unseen data. The significance of the graphs lies in their ability to guide model selection and hyperparameter tuning to achieve optimal performance.

IX. Cost Efficiency

Model	**Total Elapsed Time**
Logistic Regression	1.2142
Naive Bayes	1.0534
Neural Network	2.0456
KNN	14.2811s

The Total Elapsed Time column provides insights into the **overall efficiency** of each model in terms of the time taken for training, prediction.

1. **Logistic Regression**:Demonstrates relatively quick execution with a total elapsed time of 1.2142 seconds, making it efficient for both training and prediction.

2. **Naive Bayes**: Shows impressive efficiency with a total elapsed time of 1.0534 seconds. Despite the additional time taken for bias-variance tradeoff plotting, it remains competitive.

3. **Neural Network**: Requires a longer total elapsed time of 2.0456 seconds, indicating a more computationally intensive process during training and prediction.

4. **KNN**: Presents the longest total elapsed time at 14.2811 seconds, suggesting higher computational requirements, possibly due to its reliance on the entire dataset for prediction.

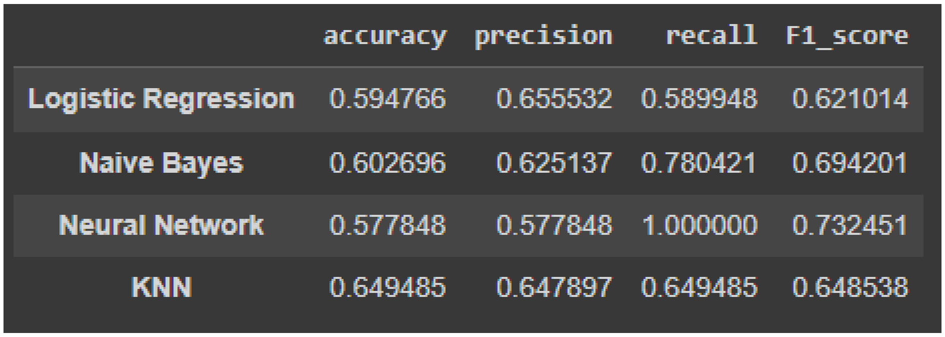
These time considerations are crucial for real-world applications, especially in scenarios where rapid predictions or resource efficiency are critical.

X. Conclusion and Summary

In the pursuit of developing an In-Vehicle Coupon Recommendation System, we delved into the complexities of **consumer behavior**, utilizing extensive data analysis and machine learning techniques. The primary objective was to **predict coupon redemption** by leveraging various contextual factors.

Four distinct models - Logistic Regression, Naive Bayes, Neural Network, and K-Nearest Neighbors (KNN) - were evaluated based on their accuracy and the crucial bias-variance trade-off.

a. Performance Metrics



	accuracy	precision	recall	F1_score
Logistic Regression	0.594766	0.655532	0.589948	0.621014
Naive Bayes	0.602696	0.625137	0.780421	0.694201
Neural Network	0.577848	0.577848	1.000000	0.732451
KNN	0.649485	0.647897	0.649485	0.648538

1. Logistic Regression

Achieved an **accuracy** of **59.5%** on the test dataset. Its **simplicity** and **interpretability** make it an attractive choice for scenarios where understanding the importance of each feature is essential.

2. Naive Bayes

Demonstrated an **accuracy** of **60.3%** on the test set. Known for its efficiency in handling categorical data, Naive Bayes provided competitive results.

3. Neural Network

Stood out with the highest **accuracy** of **57.8%** on the test data. Its capacity to capture complex, non-linear relationships proved beneficial for the intricate nature of the dataset.

4. K-Nearest neighbors (KNN)

Displayed competitive accuracy, achieving **64.99%** on the test set. KNN's simplicity and ability to capture local patterns were evident in its performance.

b. Bias-Variance Consideration

1. Logistic Regression:

- Shows signs of overfitting, and the test error starts increasing after a certain number of iterations.

2. Naive Bayes:

- Demonstrates the trade-off between model complexity and bias/variance by adjusting the number of features.

3. Neural Network:

- Indicates overfitting as the test error increases after a certain number of epochs, highlighting the need for regularization or tuning.

4. KNN:

- Shows decreasing errors as the number of neighbors increases, suggesting potential overfitting due to increased model complexity.

c. Model Comparison

1. **Logistic Regression**
   - **Performance Metrics**: **F1 score -** 0.62 is a balance between precision and recall.
   - **Bias-Variance Tradeoff**: Potential overfitting as test error starts increasing after a certain number of iterations.

2. **Naive Bayes**
   - **Performance Metrics**: **F1 score -** 0.69 is higher than Logistic Regression.
   - **Bias-Variance Tradeoff**: Increasing the number of features increases complexity but also increases the variance. The model may be prone to overfitting.

3. **Neural Network**
   - **Performance Metrics**: **F1 score -** 0.73 is the highest among the models.
   - **Bias-Variance Tradeoff**: Overfitting is indicated as the test error increases after a certain number of epochs.

4. **KNN**
   - **Performance Metrics**: F1 score **-** 0.64 is competitive but slightly lower than Naive Bayes and Neural Network.
   - **Bias-Variance Tradeoff**: Increasing the number of neighbors increases complexity but also the variance.

Thus, **Neural Network** seems to be the most suitable model for the classification problem:

- It has the **highest F1 score**, indicating a good **balance between precision and recall**.
- Although there is **overfitting**, the alignment between training and testing errors suggests a well-balanced model.
- The dataset may benefit from the non-linear capabilities of a neural network, capturing complex relationships.

XI. Limitations and Scope

a. Limitations

1. Data Constraints

- The **moderate size** of the dataset (12,684 entries) may limit the model's ability to generalize, especially for complex patterns.
- Imbalanced class distribution could pose challenges in accurately predicting the minority class, affecting model performance.

2. Feature and Model Assumptions

- Feature space limitations may overlook crucial aspects of consumer behavior, impacting predictive power.
- Assumptions of chosen models (e.g., linearity in Logistic Regression, independence in Naive Bayes) might not fully align with the data distribution.

3. Interpretability and Temporal Dynamics

- Model interpretability varies; Neural Networks may lack transparency, impacting user acceptance. The **absence of a temporal dimension** in the dataset may overlook evolving consumer trends.

b. Future Scope

1. **Advanced Modeling Techniques**

- Explore ensemble models (e.g., Random Forests, Gradient Boosting) to enhance generalization and robustness.
- Consider sophisticated deep learning architectures to capture intricate relationships within the data.

2. Enhanced Feature Engineering

- Investigate additional features from external sources to enrich the understanding of consumer behavior.

Addressing these limitations and pursuing these future directions can elevate the In-Vehicle Coupon Recommendation System, making it more accurate, adaptive, and aligned with the complexities of consumer behavior in the dining industry.

c. References

1. UCI Machine Learning Repository
2. Science Direct for machine learning models

Group 11	
Team Member	**NUID**
Achala Shenoy	002736738
Ankita Shukla	002920460
Navya Pragathi Molugaram	002774072
Sanidhya Mathur	002766999