EXPERIMENT REPORT

Name	Monali Patil
Project Name	Customer-Vehicle-Repurchase-Prediction
Deliverables	Model name: K-NearestNeighbors (KNN) Classifier Notebook name: MachineLearning_K-NearestNeighbors_Exp1.ipynb Project Repo: https://github.com/monalippatil/MachineLearning-Customer-Vehicle-Repurchase-Prediction.git

---

1. EXPERIMENT BACKGROUND

Provide information about the problem/project such as the scope, the overall objective, expectations. Lay down the goal of this experiment and what are the insights, answers you want to gain or level of performance you are expecting to reach.

1.a. Business Objective

Explain clearly what is the goal of this project for the business. How will the results be used? What will be the impact of accurate or incorrect results?

Objective: The aim of this project is to develop a binary classifier model that can predict customers who are likely to purchase a new vehicle. To achieve this, a dataset of car repurchases will be used for this binary classification problem and through data exploration, cleaning, feature scaling, selection of appropriate algorithm, hyperparameters and employing various techniques such as cross-validation, hyperparameter tuning, feature importance etc., to improve the accuracy of the model.

Application: The outcomes of this model can be used to identify the customers who are most likely to buy a new vehicle, and this information can be used to focus marketing efforts on those potential customers which can lead to an effective and efficient promotional campaign.

Impact: The model's accurate results can help the company to target the right customers who are more likely to buy the vehicle through a marketing campaign which could save marketing costs and the campaign could generate higher sales with fewer marketing expenses. However, incorrect results of the model can result in wasted marketing efforts and resources and thus, the marketing campaign may fail to achieve the desired results and sales figures.

1.b. Hypothesis	Present the hypothesis you want to test, the question you want to answer or the insight you are seeking. Explain the reasons why you think it is worthwhile considering it, Hypothesis: The objective of this experiment is to effectively identify if present customers are likely to buy a new vehicle, and to carry out this initial experiment, the KNeighbors Classifier using Euclidean Distance algorithm has been employed for conducting and evaluating the performance to determine if this algorithm is suitable for this Binary Classification task. Rationale: Considering the business objective of determining potential customers leads for a marketing campaign, there is no good theory to map and select a suitable algorithm for this binary classification problem, therefore performing different experiments to discover which algorithm and algorithm configuration results in the best performance for this binary classification task.
1.c. Experiment Objective	Detail what will be the expected outcome of the experiment. If possible, estimate the goal you are expecting. List the possible scenarios resulting from this experiment. The expected outcome of this experiment would be the development of a binary classifier model utilizing KNeighbors Classifier algorithm that can accurately predict which customers are more likely to buy a new vehicle. The model would be trained on a dataset of current customer information and their buying behavior, and it would use KNeighbors Classifier algorithm to identify patterns and make predictions about future buying behaviour of the customers. The goal of this experiment would be to identify potential customers for a promotional marketing campaign aimed at increasing vehicle sales, employing KNeighbors Classifier algorithm. By focusing on customers who are more likely to buy a new vehicle, the marketing campaign could be more targeted and effective. The possible scenarios resulting from this experiment include following: • The model accurately predicts which customers are likely to buy a new vehicle, allowing the marketing team to focus on these potential customers and increase sales. • The model has a high false positive rate, indicating that the model predicts some customers will purchase a new vehicle when they actually won't purchase. This could result in wasted marketing resources and a lower return on investment. • The model has a high false negative rate, indicating that the model fails to predict some customers will purchase a new vehicle when they actually will. This could result in missed opportunities for sales and revenue. • The model is not accurate enough to be useful for predicting customer behavior, and the experiment is unsuccessful. In this case, the marketing team may need to explore other methods for identifying potential customers for their promotional campaign.

2. EXPERIMENT DETAILS

Elaborate on the approach taken for this experiment. List the different steps/techniques used and explain the rationale for choosing them.

2.a. Data Preparation

Describe the steps taken for preparing the data (if any). Explain the rationale why you had to perform these steps. List also the steps you decided to not execute and the reasoning behind it. Highlight any step that may potentially be important for future experiments.

To prepare for using the binary classification algorithm, the data for current customers was explored, cleaned up and prepared through the subsequent activities.

• Data Understanding

1] Loading Data: To use and create a binary classification model, imported data from a CSV file into the pandas dataframe.

2] Exploring Data: Explored and analysed current customer's data utilizing various following pandas functionalities to understand and identify patterns to further determine potential customers who are more likely to purchase a new vehicle.

* Additionally, to ensure the quality of the data to be utilized by the model analyzed.
- Missing/null values.
- Duplicate records.
- Outliers for numerical features.
- Data distribution for various categorical and numerical features.
- Distinct values of the categorical features.

a) df.head(): Examining initial observations of the dataset.
b) df.shape(): Examining the dimension of the dataset.
c) df.columns: Inspecting features name.
d) df.info(): Inspecting the summary information of the attributes of the dataset.
e) df.describe(): Examining the statistical information of integer variables of the dataset.
f) df.describe(include='all'): Examining statistical summary information for all variables of the dataset across different data types.
g) df.isnull().sum(): Examining whether there are any null values in the dataset.
h) df.duplicated().sum(): Examining whether there are any missing values in the dataset.

• Data Preparation

3] Treating Missing Values

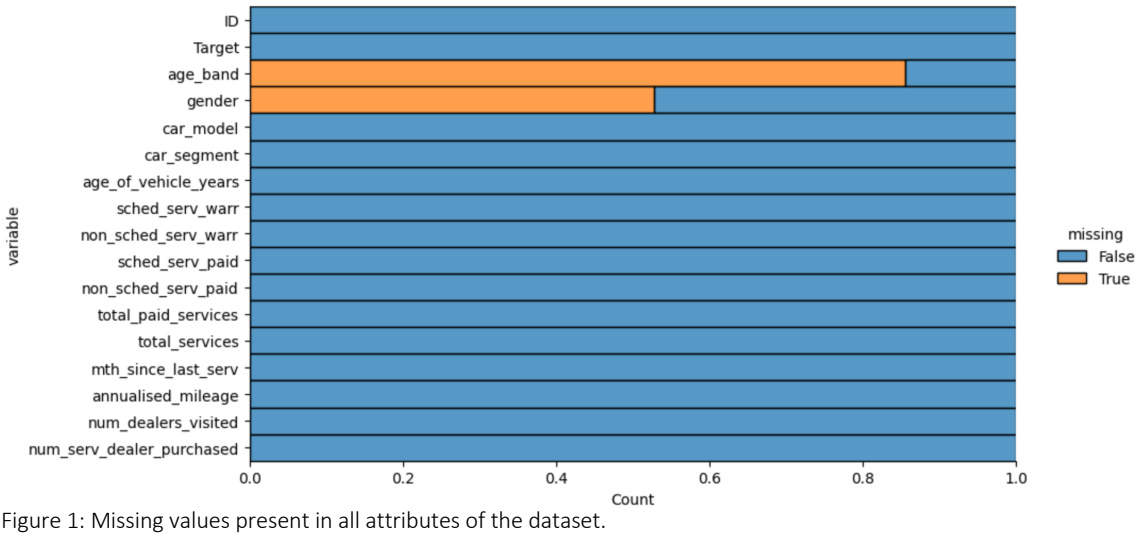

Figure 1: Missing values present in all attributes of the dataset.

* The chart from Figure 1, indicates that only two variables, 'age_band' with around 85% and 'gender' with over 50% contain missing values. However, there are no missing values for the rest of the features.

* As machine learning algorithms are incapable of processing features with missing values, we have two options to address these missing values.

i) Replace missing values with the relevant value: Using the most frequent occurring (mode) value from the feature to replace missing values in categorical variables may not accurately represent the distribution of data in those variables 'age_band' and 'gender', especially when large portion of the data is missing. This approach could also create an imbalance in the data by introducing a high number of imputed mode values, potentially leading to biased results.

ii) Eliminate the columns containing missing values entirely: It would be suitable to remove these two 'age_band' and 'gender' features since replacing any value would significantly distort the data due to the high percentage of missing values in these variables.

* Therefore, removed 'age_band' and 'gender' features from the vehicle repurchase dataset using pandas drop() function and specifying these two feature names.

4] Transforming Categorical Data into Numerical

Attribute Name	Distinct/Unique Values	Count of Distinct Values
car_model	model_1, model_2, model_3, model_5, model_6, model_4, model_7, model_8, model_9, model_10, model_11, model_13, model_12, model_14, model_15, model_16, model_17, model_18, model_19	19
car_segment	LCV, Small/Medium, Large/SUV, Other	4

Table 1: Distinct values present in ‘car_model’ and ‘car_segment’ attributes of the dataset.

* The table 1 illustrates that the features 'car_model' and 'car_segment' contain categorical data, with 19 distinct car models and 4 different vehicle segments respectively. However, this data

cannot be used directly in binary classification algorithms for machine learning because these algorithms operate on mathematical equations that require numerical inputs.

* Additionally, these features 'car_model' and 'car_segment' hold significant information regarding the model and category of vehicles that customers have purchased. This information can be used to predict whether the customer is inclined to purchase a new vehicle.

One-Hot Encoding

* Therefore, it is necessary to convert the categorical data of these features into numerical data. To achieve this, employing one-hot encoding method, which creates binary columns for every distinct category in the feature.

* This method is preferred as there is no inherent order among the category values of the feature, and do not want to create any random relationships between them and prevent model from assuming natural ordering among these distinct categories that may suffer from model bias.

5] Selecting Target Variable & Features

* The 'TARGET' feature is the target variable representing customers who have bought more than one vehicle with class 1 and those who have only purchased one vehicle with class 0. This, along with the predictor features, are separated into 'X' and 'y' variables to build a binary classification model.

6] Splitting Data into Different Sets: Training, Validation and Testing

* Creating a validation set provides adaptability to carry out multiple experiments and allows the comparison of the model's performance on the training set and the validation set to determine if the model is overfitting or underfitting.

* Furthermore, using a validation set, we can adjust model's hyperparameters and retrain the model until it performs well on the validation set. After we have optimized the model's performance on the validation set, we can then apply the model on the test set to assess its ability to generalize to new and unseen data.

* Therefore, utilizing train_test_split() function from sklearn.model_selection module of sklearn library, the car repurchase dataset is split into training (60%), validation (20%), and testing (20%) sets to leverage the flexibility of multiple experimentation.

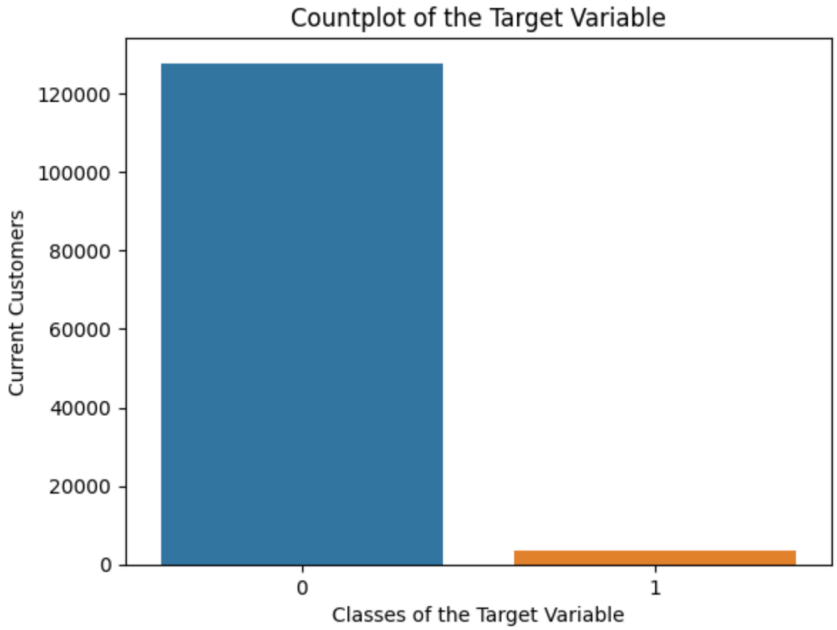

Figure 2: Examining the quantity of each class of the target variable from the actual dataset.

* Additionally, for any (binary) classification problem, it is essential to examine the frequencies/rates of every class within the target variable and ensure that the splitting of data into different sets matches as similar to the actual data spread.

* As indicated in the Figure 2, the car repurchase dataset is highly imbalanced, it is necessary to obtain as similar distribution of distinct classes of the target variable for all of the different sets as original dataset. After splitting process, following are the frequencies of the target variable classes.

Dataset/Frequency Rates	Class 1	Class 0
Actual Data:	0.97	0.026
Training Set:	0.97	0.026
Validation Set:	0.97	0.026
Testing Set:	0.97	0.028

7] Features Scaling

* Machine learning algorithms often use some form of distance calculation to determine the relationship between different features in the dataset. If the features are on different scales, the algorithm may prioritize some high value range features over others which may have more significant information, leading to biased results.

* In addition, the values for the numerical features are divided into ten deciles, ranging from 1 to 10, while categorical data that has been converted into numerical features, consist of either 0 or 1.

* Therefore, performing scaling for all the features so that all the features values are at same level and that the algorithm can use all the information from the dataset features to learn generalised patterns, identify buying behaviour and make accurate predictions.

StandardScaler

* Employing StandardScaler class which is imported from sklearn.preprocessing module to scale and bring all features values at same level. Choosing this method because it does not change the shape of the data distribution and preserves outliers as it scales the data based on the mean of 0 and standard deviation of 1, of the entire dataset, rather than individual datapoints.

The following are some crucial steps that could be important for any of the classification experiments in future.

- It is important to thoroughly examine and handle missing values in order to avoid introducing biases in the model.
- Outliers and duplicate values should be carefully analysed and reviewed from business perspectives and treated accordingly.
- The decision to use LabelEncoder, OrdinalEncoder or OneHotEncoder for converting categorical data into numerical form should depend on various factors, including the nature of the categorical values and whether or not there is any inherent order among them. Careful consideration is necessary to determine the appropriate method.
- Similarly, the selection of feature scaling method - whether to use MinMaxScaler, MaxAbsScaler, or StandardScaler - should be made based on whether or not there is an inherent order in the data, as well as other factors such as the need to maintain the integrity of outliers.
- For classification problems and especially for imbalanced data, is essential to ensure the frequency of each class in the target variable in each set is representative of the actual data. The "stratify=y" parameter could be used during the splitting process if the class distribution is not similar to the actual data. This will help to maintain the integrity of the data and ensure accurate results.
- To prevent the model from overfitting on specific data points and to enable it to learn generalized patterns, it is important to eliminate any identifier attributes.

2.b. Feature Engineering

Describe the steps taken for generating features (if any). Explain the rationale why you had to perform these steps. List also the feature you decided to remove and the reasoning behind it. Highlight any feature that may potentially be important for future experiments

	ID
count	131337.000000
mean	77097.384180
std	44501.636704
min	1.000000
25%	38563.000000
50%	77132.000000
75%	115668.000000
max	154139.000000

Figure 3: Statistical summary of 'ID' attribute.

	* The 'ID' attribute seems to be an identifier and its values ranges from 1 to 154139 as indicated in Figure 3. It does not have any predictive value as it contains a unique value for each observation, which means that its values do not contribute to any patterns or trends in the data. * And including them in the analysis can lead to overfitting straight a way, where model would fit these specific values from the feature rather than the underlying generic patterns in the data. * Therefore, removed 'ID' attribute from the vehicle repurchase dataset using pandas drop() function and specifying this feature name. Note: Removal of 'age_band' and 'gender' features is stated in earlier part 3] Treating Missing Values of 2.a Data Preparation section.
2.c. Modelling	Describe the model(s) trained for this experiment and why you choose them. List the hyperparameter tuned and the values tested and also the rationale why you choose them. List also the models you decided to not train and the reasoning behind it. Highlight any model or hyperparameter that may potentially be important for future experiments The model trained for this experiment is one of the Binary Classification algorithms, KNeighbors Classifier model using Euclidean Distance to effectively predict customers who are likely to purchase a new vehicle. The model is trained, and the performance is evaluated to determine if this algorithm is suitable for the Binary Classification task which can then be used to target potential customers for a marketing campaign. The target variable consists of either 0 or 1, where 0 represents the purchase of a single vehicle by the customer, while 1 represents the purchase of multiple vehicles. As we aim to predict only these two possible classes outcome, this Binary Classification model is the right choice along with learning objective purpose. Hyperparameters Selected. Hyperparameters Name | Values Tested n_neighbors: 15 | 7 | 3 metric: euclidean * To begin with, applied n_neighbors=15, metric='euclidean' hyperparameters to the KNN classifier algorithm and assessed its performance. Based on this evaluation, adjusting the n_neighbors hyperparameter values as needed for different exercises to achieve optimal performance. Cross-Validation: cross_val_score() * To determine the optimal value of 'k' neighbors to ensure that it aligns with the model that had a hyperparameter value of n_neighbors=3 and had achieved ideal performance scores based on all the trained models in this experiment.

* Employed cross-validation to identify ideal value for 'k' because the cross-validation performs training and testing of the model over multiple folds of the dataset to promote model generalisation and to understand the model performance over the whole dataset rather than relying on a single train-validation split.

* Therefore, for this, utilized cross_val_score() cross validation function for different values of k (ranging from 1 to 30) in the KNeighbors classifier and determined the value of k that gives the highest cross-validation score.

Due to time constraints, it was not possible to test the KNN classifier algorithm with additional hyperparameters like algorithm and leaf_size. However, I plan to test these hyperparameters in future experiments.

3. EXPERIMENT RESULTS

Analyse in detail the results achieved from this experiment from a technical and business perspective. Not only report performance metrics results but also any interpretation on model features, incorrect results, risks identified.

3.a. Technical Performance

Score of the relevant performance metric(s). Provide analysis on the main underperforming cases/observations and potential root causes.

Performance Metrics

To evaluate the model's performance utilizing the below performance metrics.

- Precision
- Recall
- Weighted F1 Score
- Binary F1 Score
- Confusion Matrix

Baseline Performance (Null Accuracy)

weighted F1 Score: 0.9608

Exercise 1:

Algorithm: KNeighbors (KNN) Classifier model

Hyperparameters: n_neighbors=15, metric='euclidean'

Performance Metrics:

Dataset	Precision	Recall	weighted F1 Score	binary F1 Score
Training:	0.9425	0.4165	0.9809	0.5778
Validation:	0.9288	0.3705	0.9786	0.5297

	Error Rate		Error Count	
Dataset	False Positive	False Negative	False Positive	False Negative
Training:	0.07%	1.53%	56	1287
Validation:	0.08%	1.69%	16	355

* The binary F1 score resulting positive label classes of the validation set (0.5297) is relatively lower than that of the training set (0.5778), and the weighted F1 score of the validation set (0.9786) is slightly lower than training set (0.9809), suggesting that the model is moderately overfitting. s

* Moreover, the difference between the precision (and recall) scores from the training and validation set informs the same. Also, the False Negative error from the confusion matrix of both the training set with 1287 and the validation set with 355 are significant.

* Therefore, attempting to address the issue of overfitting and lowering the False Negative error with different value of 'n_neighbors' hyperparameter and thus, adjusting the hyperparameter n_neighbors value to 7.

Exercise 2:

Algorithm: KNeighbors (KNN) Classifier model

Hyperparameters: n_neighbors=7, metric='euclidean'

Performance Metrics:

Dataset	Precision	Recall	weighted F1 Score	binary F1 Score
Training:	0.9485	0.5344	0.9852	0.6836
Validation:	0.9033	0.4308	0.9806	0.5834

	Error Rate		Error Count	
Dataset	False Positive	False Negative	False Positive	False Negative
Training:	0.08%	1.22%	64	1027
Validation:	0.12%	1.53%	26	321

* Continually, the binary F1 score of the validation set (0.5834) is relatively lower than the training set (0.6836) and similarly, the difference for the Precision and Recall score, informs the model is overfitting. Although the overall performance of the model with the weighted F1 score seems fine.

* Moreover, the False Negative error from the confusion matrix has slightly decreased as compared to the earlier model but is still substantial with 1027 for training and 321 for the validation set, indicating that the model is misclassifying existing customers that are more likely to purchase a new vehicle and thus, requires further investigation to reduce it.

* Consequently, setting the value of the hyperparameter n_neighbors to 3 based on the learning from the above exercise and its evaluation results.

Exercise 3:

Algorithm: KNeighbors (KNN) Classifier model

Hyperparameters: n_neighbors=3, metric='euclidean'

Performance Metrics:

Dataset	Precision	Recall	weighted F1 Score	binary F1 Score
Training:	0.9610	0.6817	0.9901	0.7976
Validation:	0.8629	0.5248	0.9832	0.6527
Testing:	0.8750	0.5126	0.9819	0.6465

	Error Rate		Error Count	
Dataset	False Positive	False Negative	False Positive	False Negative
Training:	0.07%	0.84%	61	702
Validation:	0.22%	1.28%	47	268
Testing:	0.21%	1.39%	55	366

* The binary F1 score of the training set (0.7976) is relatively higher than the validation (0.6527) and testing set (0.6465) indicating that even the well-chosen hyperparameter value n_neighbors=3 (obtained by utilizing cross_val_score() cross validation function) with Euclidean distance, the KNN classifier model is significantly overfitting as some of the datapoints might be too specific and not generic enough. The Precision and Recall measures also illustrate the same overfit nature of the model. While the weighted F1 score for the three sets suggests slight overfitting.

* Additionally, the False Negative errors of the training set with 702, validation with 268 and testing with 366 are considerable and need to be addressed to ensure effective and efficient use of company resources for the marketing campaign.

* So even though the model is able to significantly identify the customers who are likely to buy a new vehicle, it has not generalised well on unseen test data. Therefore, considering to explore other classification algorithms, Logistic Regression to thoroughly evaluate and select the most suitable model for identifying potential customers for the marketing campaign that are willing to purchase a new vehicle.

3.b. Business Impact

Interpret the results of the experiments related to the business objective set earlier. Estimate the impacts of the incorrect results for the business (some results may have more impact compared to others)

All the 3 binary classification exercises, employing KNeighbors (KNN) Classifier model with different n_neighbors (15, 7 and 3) values, is not sufficiently generalized and there are still features with observations that are specific, leading to overfitting. Thus, the model is not accurate enough to be useful for predicting customer who are likely to purchase a new vehicle.

Moreover, there are considerable amount of False Negative errors, indicating that the model is misclassifying customer who are likely to purchase a new vehicle as not willing to purchase leading to missed opportunities for sales and revenue and potentially resulting in the failure to meet marketing campaign objectives or targets.

Therefore, it is necessary to address this issue from a business perspective, considering marketing

	campaign goals and benchmarks to be achieved.
3.c. Encountered Issues	List all the issues you faced during the experiments (solved and unsolved). Present solutions or workarounds for overcoming them. Highlight also the issues that may have to be dealt with in future experiments. Due to a large number of missing values in the categorical features 'age_band' and 'gender', there was uncertainty regarding whether to replace the missing values with the mode value or remove the features altogether. However, after studying several articles and considering the business problem at hand, decided to remove these features would be a suitable choice. When it came to converting categorical data to numerical and scaling the features, there was confusion about which method to use. Therefore, studied several articles, examples and conducted small experiments to gain a better understanding of the different methods and their outcomes. Ultimately, after careful consideration of feature values and the problem, chose OneHotEncoder and StandardScaler method. As the performance metric needed to be selected, the process of selecting appropriate performance metric involved extensive analysis of lab sessions, canvas materials, several articles. Additionally, discussed with the Professor to clarify doubts and then chose the performance metric mentioned in the 3.a Technical Performance section.

---

4. FUTURE EXPERIMENT	
Reflect on the experiment and highlight the key information/insights you gained from it that are valuable for the overall project objectives from a technical and business perspective.	
4.a. Key Learning	Reflect on the outcome of the experiment and list the new insights you gained from it. Provide rationale for pursuing more experimentation with the current approach or call out if you think it is a dead end. Despite the KNeighbors (KNN) Classifier model being trained using the ideal 'n_neighbors' hyperparameter value of 3, which was derived through cross-validation using the cross_val_score() function, the model was still significantly overfitting. Furthermore, the model produced a considerable number of False Negative errors. Therefore, even though the optimal 'k' neighbors value was used, it does not guarantee the best performance of the model. Additional analysis is required, such as cross-validation and exploration of other hyperparameters that can be penalized, to ensure that the model can be more generalized. Moreover, it is crucial to consider the business context and perspective when performing machine learning experiments to evaluate model performance and determine if it is worth continuing with further analysis.

	As stated in the 2.c Modelling section, there is an opportunity for further evaluation and analysis. Therefore, in future experiments, I plan to test additional hyperparameters of the KNN classifier to overcome the level of overfitting in the model and enhance its ability to generalize.
4.b. Suggestions / Recommendations	Given the results achieved and the overall objective of the project, list the potential next steps and experiments. For each of them assess the expected uplift or gains and rank them accordingly. If the experiment achieved the required outcome for the business, recommend the steps to deploy this solution into production. Primarily would be performing these two tasks. * Train the KNN classifier algorithm with additional hyperparameters like algorithm and leaf_size and review if the performance is improving. * Analyse the model's performance using ROC curves and Precision Recall Curves. To successfully deploy a binary classification algorithm in a production environment, it is advisable to carry out below steps. - Scale and transform the model to handle large datasets. - Select an appropriate deployment environment whether cloud-based or on-premises. - Modify the model to suit production settings while complying with various security, ethical, and privacy guidelines. - Conduct testing and monitoring of the deployed model. - Periodically updated the model with new data. - Provided documentation for usage. - Review the performance of the model and retrain the model as needed.