Project Support Session

Agenda

- Machine Learning Quiz
- Personal Loan Campaign - Problem Statement
- Personal Loan Campaign - Dataset
- Personal Loan Campaign - Project FAQs
- QnA

Let's begin the discussion by answering a few questions on machine learning

Machine Learning Quiz

In linear regression, what is the purpose of the R-squared metric?

A. To measure the strength and direction of the linear relationship between independent and dependent variables

B. To assess the contribution of individual coefficients in the regression model

C. To evaluate the how well the model explains the variability in the data

D. To identify the value of the dependent variables when all the independent variables are zero

Machine Learning Quiz

In linear regression, what is the purpose of the R-squared metric?

A. To measure the strength and direction of the linear relationship between independent and dependent variables

B. To assess the contribution of individual coefficients in the regression model

C. To evaluate the how well the model explains the variability in the data

D. To identify the value of the dependent variables when all the independent variables are zero

Machine Learning Quiz

In linear regression, R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. In other words, it quantifies how well the regression model fits the observed data. R-squared can take a maximum value of 1, where 1 indicates a perfect fit.

Machine Learning Quiz

A very complex model will more likely perform better on the test data set than the train data set.

A. True

B. False

Machine Learning Quiz

A very complex model will more likely perform better on the test data set than the train data set.

A. True

B. False

Machine Learning Quiz

A highly complex model is prone to overfitting the training data by capturing noise rather than general patterns. Consequently, it is likely to perform worse on unseen data, such as the test dataset, compared to its performance on the training dataset.

Machine Learning Quiz

While building a multiple linear regression model, it was found that the addition of a variable decreased the value of the adjusted R-squared.

Which of the following statements is correct?

A. The new variable should be added to the final model

B. The new variable should not be added to the final model

Machine Learning Quiz

While building a multiple linear regression model, it was found that the addition of a variable decreased the value of the adjusted R-squared.

Which of the following statements is correct?

A. The new variable should be added to the final model

B. The new variable should not be added to the final model

Machine Learning Quiz

If the addition of a variable decreases the value of the adjusted R-squared, it indicates that the variable does not contribute significantly to explaining the variation in the target variable after accounting for the existing variables in the model. Therefore, adding this variable to the final model could lead to overfitting and reduced predictive performance. It's generally recommended to exclude variables that do not improve the model's performance or contribute meaningfully to the understanding of the target variable.

Machine Learning Quiz

Great Learning
POWER AHEAD

It is necessary to create dummies for columns that only have 0's and 1's.

A Yes

B No

Machine Learning Quiz

Great Learning
POWER AHEAD

It is necessary to create dummies for columns that only have 0's and 1's.

A Yes

B No

Machine Learning Quiz

It is not necessary to create dummy variables for columns that only contain binary values (0's and 1's). Since these columns already represent categorical information in a binary format, creating dummy variables would be redundant.

Machine Learning Quiz

Great Learning
POWER AHEAD

Which of the following methods is used to detect outliers present in the data?

A. One hot encoding

B. Interquartile range (IQR) method

C. Confusion Matrix

D. MSE

Machine Learning Quiz

Great Learning
POWER AHEAD

Which of the following methods is used to detect outliers present in the data?

A. One hot encoding

B. Interquartile range (IQR) method

C. Confusion Matrix

D. MSE

Machine Learning Quiz

One hot encoding	One-hot encoding creates binary variables (dummy variables) from categorical variables.
Interquartile range (IQR) method	The Interquartile range (IQR) method is a statistical technique used to identify outliers in data by measuring the dispersion of the dataset.
Confusion Matrix	Confusion matrix is a technique used to evaluate the performance of classification models
MSE	Mean Squared Error (MSE) is a metric used to measure the average squared difference between the actual and predicted values in regression models

Machine Learning Quiz

Great Learning
POWER AHEAD

Which of the following parameters is used to control the maximum depth of the decision tree?

A random_state

B min_samples_split

C max_depth

D min_samples_leaf

Machine Learning Quiz

Great Learning
POWER AHEAD

Which of the following parameters is used to control the maximum depth of the decision tree?

A. random_state

B. min_samples_split

C. max_depth

D. min_samples_leaf

Machine Learning Quiz

random_state	random_state is used to assign a fixed seed value to control randomness and enable reproducible results
min_samples_split	min_samples_split is used to specify the minimum number of samples required to split an internal node
max_depth	max_depth is used to specify the maximum depth of the decision tree, controlling how deep the tree can grow.
min_samples_leaf	min_samples_leaf is used to specify the minimum number of samples required to be at a leaf node

Machine Learning Quiz

What is the primary purpose of pruning in decision tree?

A. To reduce model complexity and overfitting

B. To increase the size of the model for better performance

C. To speed up the training process

D. To introduce noise into the model for better generalization

Machine Learning Quiz

What is the primary purpose of pruning in decision tree?

A. To reduce model complexity and overfitting

B. To increase the size of the model for better performance

C. To speed up the training process

D. To introduce noise into the model for better generalization

Machine Learning Quiz

Pruning in decision trees involves removing unnecessary branches or nodes to simplify the model. The primary aim is to reduce model complexity and prevent overfitting, where the model fits too closely to the training data and performs poorly on new data. By pruning, decision trees become more generalized, capturing essential patterns in the data without memorizing noise, thus improving their ability to make accurate predictions on unseen data.

Machine Learning Quiz

Which of the statement describes pre-pruning and post-pruning in decision tree?

A. Pre-pruning removes unnecessary branches from a fully grown tree; post-pruning stops tree growth early based on criteria

B. Pre-pruning stops tree growth early based on criteria; post-pruning removes unnecessary branches from a fully grown tree

C. Both pre-pruning and post-pruning involve halting tree growth early based on criteria

D. Both pre-pruning and post-pruning involve removing unnecessary branches from a fully grown tree

Machine Learning Quiz

Which of the statement describes pre-pruning and post-pruning in decision tree?

A. Pre-pruning removes unnecessary branches from a fully grown tree; post-pruning stops tree growth early based on criteria

B. Pre-pruning stops tree growth early based on criteria; post-pruning removes unnecessary branches from a fully grown tree

C. Both pre-pruning and post-pruning involve halting tree growth early based on criteria

D. Both pre-pruning and post-pruning involve removing unnecessary branches from a fully grown tree

Machine Learning Quiz

Pre-pruning involves stopping the tree construction process prematurely based on certain conditions, such as reaching a maximum depth or minimum number of samples in a node, to prevent overfitting. Post-pruning occurs after the tree is fully grown, where sub-tree branches are removed from the tree.

Machine Learning Quiz

Great Learning
POWER AHEAD

Suppose you're working on a medical diagnosis task where correctly identifying patients with a certain disease is crucial. Which classification metric would you prioritize to minimize the chances of missing positive cases?

A. Accuracy

B. Precision

C. Recall

D. F1 Score

Machine Learning Quiz

Great Learning
POWER AHEAD

Suppose you're working on a medical diagnosis task where correctly identifying patients with a certain disease is crucial. Which classification metric would you prioritize to minimize the chances of missing positive cases?

A. Accuracy

B. Precision

C. Recall

D. F1 Score

Machine Learning Quiz

In a medical screening test for a rare disease, you want to ensure that the test correctly identifies all individuals with the disease. Recall would be the appropriate metric as it measures the proportion of true positive cases (individuals with the disease) that are correctly identified by the test.

Machine Learning Quiz

Great Learning
POWER AHEAD

Gini impurity ranges from 0 to 0.5, where a value closer to 0 indicates a pure node (all samples belong to the same class), and a value closer to 0.5 indicates maximum impurity (samples are evenly distributed across different classes).

A. True

B. False

Machine Learning Quiz

Great Learning
POWER AHEAD

Gini impurity ranges from 0 to 0.5, where a value closer to 0 indicates a pure node (all samples belong to the same class), and a value closer to 0.5 indicates maximum impurity (samples are evenly distributed across different classes).

A True

B False

Machine Learning Quiz

Great Learning
POWER AHEAD

Gini impurity is a measure of impurity or uncertainty in a set of data. It ranges from 0 to 0.5, where a Gini impurity of 0 indicates perfect purity (all samples belong to the same class), and a Gini impurity of 0.5 indicates maximum impurity (samples are evenly distributed across different classes). Therefore, a lower Gini impurity score corresponds to a more homogeneous set of samples, while a higher Gini impurity score indicates greater diversity or mixing of samples across different classes.

Machine Learning Quiz

What is the primary objective of K-means Clustering?

A. To minimize the within-cluster sum of squares

B. To maximize the between-cluster sum of squares

C. To minimize the silhouette score

D. To maximize the MSE

Machine Learning Quiz

What is the primary objective of K-means Clustering?

A. To minimize the within-cluster sum of squares

B. To maximize the between-cluster sum of squares

C. To minimize the silhouette score

D. To maximize the MSE

Machine Learning Quiz

The primary objective of the K-means clustering algorithm is to partition a dataset into K clusters in such a way that the within-cluster sum of squares (WCSS) is minimized. This means that the algorithm aims to minimize the distance between data points within the same cluster, resulting in tight, compact clusters.

Machine Learning Quiz

Great Learning
POWER AHEAD

Scaling is very important in machine learning because all features do not have the same range of values, and scaling prevents any feature from becoming dominant due to the high magnitude of its values.

A. True

B. False

Machine Learning Quiz

Great Learning
POWER AHEAD

Scaling is very important in machine learning because all features do not have the same range of values, and scaling prevents any feature from becoming dominant due to the high magnitude of its values.

A. True

B. False

Machine Learning Quiz

Scaling is indeed crucial in machine learning because features often have different scales or ranges of values. Without scaling, features with larger scales may dominate the learning process, leading to biased models where certain features have more influence than others solely due to their magnitude. Scaling helps ensure that all features contribute equally to the model's learning process by bringing them to a comparable scale.

Machine Learning Project

Personal Loan Campaign - Problem Statement

- AllLife Bank is a US bank that has a growing customer base. The majority of these customers are liability customers (depositors) with varying sizes of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors).
- A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns with better target marketing to increase the success ratio.
- You as a Data scientist at AllLife bank have to build a model that will help the marketing department to identify the potential customers who have a higher probability of purchasing the loan.
- To predict whether a liability customer will buy personal loans, to understand which customer attributes are most significant in driving purchases, and to identify which segment of customers to target more.

Personal Loan Campaign - Dataset

- ID: Customer ID
- Age: Customer's age in completed years
- Experience: #years of professional experience
- Income: Annual income of the customer (in thousand dollars)
- ZIP Code: Home Address ZIP code.
- Family: the Family size of the customer
- CCAvg: Average spending on credit cards per month (in thousand dollars)
- Education: Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
- Mortgage: Value of house mortgage if any. (in thousand dollars)
- Personal_Loan: Did this customer accept the personal loan offered in the last campaign?
- Securities_Account: Does the customer have securities account with the bank?
- CD_Account: Does the customer have a certificate of deposit (CD) account with the bank?
- Online: Do customers use internet banking facilities?
- CreditCard: Does the customer use a credit card issued by any other Bank (excluding All life Bank)?

Machine Learning Project FAQs

Personal Loan Campaign - Project FAQs

How should one approach the Personal Loan Campaign project?

- Before starting the project, please read the problem statement carefully and go through the criteria and descriptions mentioned in the rubric.
- Once you understand the task, download the dataset and import it into a Python notebook to get started with the project.
- To work on the project, you should start with data preprocessing and EDA using descriptive statistics and visualizations.
- Once the EDA is completed and data is preprocessed, you can use the data to build a model and check its performance.
- It is important to close the analysis with key findings and recommendations to the business.

Personal Loan Campaign - Project FAQs

I'm trying to post-prune the decision tree. But I'm getting the following error:

"***ValueError***: *ccp_alpha must be greater than or equal to 0"*

How to resolve this?

To resolve this error kindly use absolute values (positive value) of alpha. Use the following lines of code to resolve the error:

```
ccp_alphas, impurities = abs(path.ccp_alphas), path.impurities
```

Personal Loan Campaign - Project FAQs

Do I need to plot every variable in the data for univariate analysis?

Yes, univariate analysis needs to be done for each of the variables in the data, which gives insights into the individual behavior and characteristics.

Personal Loan Campaign - Project FAQs

How many models do we need to build in this project?

You need to build three models for this project and below are the one

1. Decision Tree - Base Model
2. Decision Tree - Pre-Pruning
3. Decision Tree - Post-Pruning

Machine Learning Project Submission

Personal Loan Campaign - Low-code Version

- For learners who aspire to be in managerial roles in the future, focusing on solution review, interpretation, recommendations, and communication with business stakeholders
- Steps Involved
    - Download the dataset () and the *Learner Notebook - Low Code*, (this is a template notebook)
    - Fill in the blanks in the notebook to complete and execute the code to solve the questions and perform all the tasks as per the grading rubric
    - Once the notebook is completely executed and necessary outputs obtained, a business presentation (using Microsoft PowerPoint, Google Slides, etc.) has to be created
    - The presentation should contain observations, insights, and recommendations for the business problem
        - The presentation template provided can be referred to as a sample
    - Once the presentation is complete, convert the presentation to .pdf format
- **The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file**
- **Please make sure that all the sections mentioned in the grading rubric have been covered in the submission**

Personal Loan Campaign - Full-code Version

- For learners who aspire to be in hands-on coding roles in the future, focusing on building solution codes from scratch
- Steps Involved:
    - Download the dataset and the Learner Notebook - Full Code (this is a template notebook containing high-level steps to perform and insight-based questions)
    - Write necessary code to solve the questions and perform all the tasks as per the grading rubric
    - Clearly write down observations, insights, and recommendations for the business problem based on the analysis performed
    - Once the notebook is complete, download it as a .ipynb file and convert it to a .html file
- The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb)
    - The conversion can be done via one of the following ways:
        - Jupyter Notebook:
        - Google Colab: Use free online tools
- Please make sure that all the sections mentioned in the grading rubric have been covered in the submission

Personal Loan Campaign - QnA

Great Learning

Happy Learning !