|  |  |  |
| --- | --- | --- |
|  | **Use Case** | **Tasks** |
|  | -----------------------------------  ## **Problem Statement -**  #### **Daimler challenge - reduce the time that cars spend on the test bench.**  -----------------------------------  - The objective of the Mercedes-Benz Greener Manufacturing competition is to develop a machine learning model that can      - accurately `predict the time a car will spend on the test bench` based on the vehicle configuration.      - `reduce the total time spent testing vehicles by allowing cars with similar testing configurations to be run successively.  - The `vehicle configuration` is defined as the set of customization options and features selected for the particular vehicle.  - This problem is an example of a machine learning `regression` task because it requires predicting a continuous target variable (the duration of the test) based on one or more explanatory variables (the configuration of the vehicle).  - This problem is also a supervised task because the targets for the training data are known ahead of time and the model will learn based on labeled data.      2. Prepare the data to feed into the machine learning model.      3. Select an appropriate algorithm/method for the problem.      4. Optimize the model using the labeled training data.  location\_train = r"D:\AI-DATASETS\01-MISC\merc-train.csv" (5000 rows)  location\_test  = r"D:\AI-DATASETS\01-MISC\merc-test.csv" (ignored) | * **Understanding of data** * **Nulls** * **Duplicates** * **Types of data columns** * **Analyze the cardinality** * **Histogram/distribution of y column** * **Apply Filter methods of Feature selection (var, corr)** * **Apply Inferential statistics to check the usefulness of categorical columns** * **List the columns to be dropped** * **On the cleaned data, perform KNN and Linear regression** * **Observe the regression performance metrics** * **Save the model/ load/predict** |
|  | **Credit score – classification**   | **Field** | **Description** | | --- | --- | | **ID** | Unique ID of the record | | **Customer\_ID** | Unique ID of the customer | | **Month** | Month of the year | | **Name** | The name of the person | | **Age** | The age of the person | | **SSN** | Social Security Number of the person | | **Occupation** | The occupation of the person | | **Annual\_Income** | The Annual Income of the person | | **Monthly\_Inhand\_Salary** | Monthly in-hand salary of the person | | **Num\_Bank\_Accounts** | The number of bank accounts of the person | | **Num\_Credit\_Card** | Number of credit cards the person is having | | **Interest\_Rate** | The interest rate on the credit card of the person | | **Num\_of\_Loan** | The number of loans taken by the person from the bank | | **Type\_of\_Loan** | The types of loans taken by the person from the bank | | **Delay\_from\_due\_date** | The average number of days delayed by the person from the date of payment | | **Num\_of\_Delayed\_Payment** | Number of payments delayed by the person | | **Changed\_Credit\_Card** | The percentage change in the credit card limit of the person | | **Num\_Credit\_Inquiries** | The number of credit card inquiries by the person | | **Credit\_Mix** | Classification of Credit Mix of the customer | | **Outstanding\_Debt** | The outstanding balance of the person | | **Credit\_Utilization\_Ratio** | The credit utilization ratio of the credit card of the customer | | **Credit\_History\_Age** | The age of the credit history of the person | | **Payment\_of\_Min\_Amount** | Yes if the person paid the minimum amount to be paid only, otherwise no. | | **Total\_EMI\_per\_month** | The total EMI per month of the person | | **Amount\_invested\_monthly** | The monthly amount invested by the person | | **Payment\_Behaviour** | The payment behaviour of the person | | **Monthly\_Balance** | The monthly balance left in the account of the person | | **Credit\_Score** | The credit score of the person | | * **Null values** * **Duplicates** * **Outlier assessment in numeric columns (use IQR method)** * **EDA/Viz**   + **Relation between 'Occupation' & 'Credit\_Score'**   + **Annual Income of the person impacts your credit scores or not**   + **the monthly in-hand salary impacts credit scores or not**   + **if having more bank accounts impacts credit scores or not**   + **the impact on credit scores based on the number of credit cards you have**   + **the impact on credit scores based on how much average interest you pay on loans and EMIs**   + **how many loans you can take at a time for a good credit score**   + **if delaying payments on the due date impacts your credit scores or not**   + **if frequently delaying payments will impact credit scores or not**   + **if having more debt will affect credit scores or not**   + **if having a high credit utilization ratio will affect credit scores or not**   + **how the credit history age of a person affects credit scores**   + **how many EMIs you can have in a month for a good credit score**   + **if your monthly investments affect your credit scores or not**   + **if having a low amount at the end of the month affects credit scores or not** * **Perform Inferential statistics (t-test or ANOVA, as appr) for the above EDA items** * **Perform wrapper method to understand top contributing columns (use any classifier of your choice)** * **Study the outcomes from 1) EDA 2) inferential stats 3) wrapper method** * **ML Modeling**   + **Scaling if needed**   + **Encoding (explain which method chosen)**   + **Try KNN, Dec trees, RF, Logistic Regression**   + **Metrics (accuracy, confusion matrix, classification report)**   + **Learning curve**   + **Pick the best model and perform exhaustive hyperparameter tuning for the selected model, show improvement in metrics (if possible)** |
|  | machine learning project on the credit card transactions dataset  **1. Descriptive Statistics**   * Compute summary statistics (mean, median, mode, standard deviation, etc.) for numerical features. * Analyze the distribution of numerical features using histograms or box plots.   **2. Class Imbalance**   * Visualize the class distribution to highlight the imbalance between fraudulent and non-fraudulent transactions using bar plots or pie charts.   **3. Correlation Analysis**   * Compute the correlation matrix to identify relationships between features. * Visualize the correlation matrix using a heatmap to easily spot strong correlations.   **4. Feature Distribution by Class**   * Compare the distribution of numerical features for fraudulent and non-fraudulent transactions using histograms, KDE plots, or box plots.   **5. Time-based Analysis**   * Analyze the time-based patterns in the data (e.g., transaction time, date). * Plot the number of transactions over time to identify any temporal trends or patterns. * Compare the time-based patterns for fraudulent and non-fraudulent transactions.   **6. Feature Relationships**   * Scatter plots to explore relationships between pairs of features, colored by class. * Pair plots to visualize relationships between multiple pairs of features simultaneously.   **7. Outlier Detection**   * Use box plots to identify outliers in numerical features. * Analyze the presence of outliers in both fraudulent and non-fraudulent transactions.   **8. Distribution of Transaction Amounts**   * Analyze the distribution of transaction amounts using histograms or KDE plots. * Compare the distribution of transaction amounts for fraudulent and non-fraudulent transactions.   **10. Customer Behavior Analysis**   * Analyze features related to customer behavior, such as the number of transactions, average transaction amount, etc. * Visualize customer behavior metrics for fraudulent and non-fraudulent transactions.   Model building - try the following models (KNN Classifier, Logistic Regression, DT, RF, SVM) - compare the acc, confusion matrix and classification report - try data augmentation with the best model (from above) - Next use 1-class SVM and train on majority class data and predict the minority data as anomalies (evaluate using the decision function – Provide explanation) |  |