Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

NAAC Accredited with "A" Grade (CGPA : 3.18)

Department of Computer Science and Engineering (Data Science)

Subject: Applied Data Science (DJ19DSL703)

Experiment -4

(Data Preparation Best Practice)

NAME: Rishabh Patil

SAP: 60009200056

BATCH: D12

Aim: To complete Data Preparation Best Practice.

Theory:

Data preparation is a crucial step in any data science project, and it involves various activities, including gathering suitable data, determining key performance indicators (KPIs), and creating dashboards for business stakeholders.

Writing Key Performance Indicators (KPIs) for a data science project involves defining measurable metrics that align with the project's objectives and help assess its success.

Proof of Concept (POC) refers to a demonstration or preliminary test conducted to assess the feasibility and viability of an idea, concept, or solution. It serves as an initial evaluation to determine if the idea is both possible and worth pursuing.

1. Gathering Suitable Data for a Data Science Problem:

a. Understand the Problem:

- Clearly define the problem you are trying to solve with data science. Understand the business objectives and goals.

b. Identify Data Sources:

- Determine where you can obtain the necessary data. Sources may include databases, APIs, external data providers, and internal data repositories.

c. Data Relevance and Quality:

- Ensure that the data you gather is relevant to the problem at hand. Irrelevant data can lead to noise and inefficiencies.
- Assess the quality of the data by checking for completeness, accuracy, consistency, and any missing or duplicate values.

d. Data Collection Plan:

- Develop a plan that outlines how you will collect data. This includes specifying the data sources, collection methods, frequency, and responsible individuals or teams.

e. Data Licensing and Compliance:

- Be aware of data licensing and legal considerations. Ensure that you have the rights to use the data for your intended purpose.
- Comply with data protection regulations (e.g., GDPR) if applicable.

f. Data Sampling:

- For large datasets, consider taking a representative sample for initial exploration and analysis. This can save time and resources.

2. Determine All Key Performance Indicators (KPIs):

a. Understand Business Objectives:

- Collaborate closely with business stakeholders to understand their goals and objectives. KPIs should align with these business goals.

b. Identify Relevant Metrics:

- Determine which metrics and measurements are relevant to assess the success of your project.
- Prioritize a small set of critical KPIs to focus your efforts.

c. Define Measurement Methods:

- Clearly define how each KPI will be measured. This includes specifying data sources, calculation methods, and any necessary formulas.

d. Set Baselines and Targets:

- Establish baseline values for KPIs to provide context. Determine what constitutes success by setting target values or thresholds.

e. Monitor and Report:

- Implement a system to monitor KPIs continuously. Automate data collection and reporting when possible.
- Create dashboards or reports to visualize KPI trends and share them with stakeholders.

f. Iterate and Refine:

- Be prepared to iterate on your KPIs as your project evolves. If certain KPIs are not providing actionable insights, consider refining or adding new ones.

3. Business Stakeholders POC Dashboard:

a. Identify Stakeholder Needs:

- Collaborate with business stakeholders to understand their requirements and the specific insights they need from the data.

b. Dashboard Design:

- Design the dashboard with a focus on user experience and simplicity. Ensure that it is easy to navigate and understand.
- Choose appropriate visualization types (e.g., charts, graphs, tables) based on the nature of the data and the insights you want to convey.

c. Real-Time or Periodic Updates:

- Determine whether the dashboard should provide real-time or periodic (e.g., daily, weekly) updates.

d. Data Security and Access Control:

- Implement security measures to ensure that sensitive data is protected. Define access control rules based on user roles and permissions.

e. User Training and Support:

- Provide training and support to business stakeholders to help them make the most of the dashboard.

f. Feedback and Iteration:

- Continuously gather feedback from stakeholders to improve the dashboard over time. Ensure that it remains aligned with their evolving needs.

Lab Assignment:

1. Write concise KPI statements that clearly describe what each metric measures and how it aligns with project objectives. Use specific, action-oriented language. Each KPI statement should follow the SMART criteria:

    - Specific: Clearly define what is being measured.
    - Measurable: Specify how the metric will be quantified.
    - Achievable: Ensure that targets are realistic.
    - Relevant: Ensure that the metric is directly related to project goals.
    - Time-bound: Define a timeframe for achieving the target or benchmark.

Example:

KPI: **Customer Churn Rate**
Definition: The percentage of customers who stopped purchasing from the company during a specific time period.
Measurement: (Number of customers lost in a period / Total number of customers at the beginning of the period) * 100.
Target: Reduce churn rate by 15% within the next six months.

2. Create a POC dashboard using Power BI for the sample dataset of the chosen data science project.

Solution:

Car Prediction KPI Report

Business Objectives:

The project aims to predict car prices accurately based on various features to assist in setting competitive prices and optimizing inventory.

Mean Absolute Error (MAE)

Specific: Measures the average absolute difference between the predicted car prices and the actual prices.

Measurable: Calculated as the average of the absolute differences between predicted and actual prices.

Achievable: The target MAE will be set based on the historical prediction errors and industry benchmarks.

Relevant: Reflects the accuracy of the price predictions, directly impacting the project's goal of accurate price estimation.

Time-bound: Reduce MAE by 10% within the next quarter compared to the current baseline.

R-squared (R2)

Specific: Measures the proportion of the variance in car prices that is predictable from the independent variables.

Measurable: Calculated as 1 - (the sum of squared errors divided by the total sum of squares).

Achievable: The target R2 will be set based on the desired level of predictive power and industry standards.

Relevant: Indicates the predictive power of the model, aligning with the project's objective of accurate price prediction.

Time-bound: Increase R2 to 0.8 within the next two quarters from the current baseline. Feature Importance Score

Specific: Measures the relative importance of each feature in predicting car prices. Measurable: Calculated using algorithms like Random Forest or Gradient Boosting to assign scores to each feature.

Achievable: The target score will be set based on the significance of features in price prediction.

Relevant: Helps in identifying the most influential features for accurate price prediction, aligning with the project's goal.

Time-bound: Increase the importance score of the top 3 features by 15% within the next quarter compared to the current baseline.

Monitoring and Reporting:

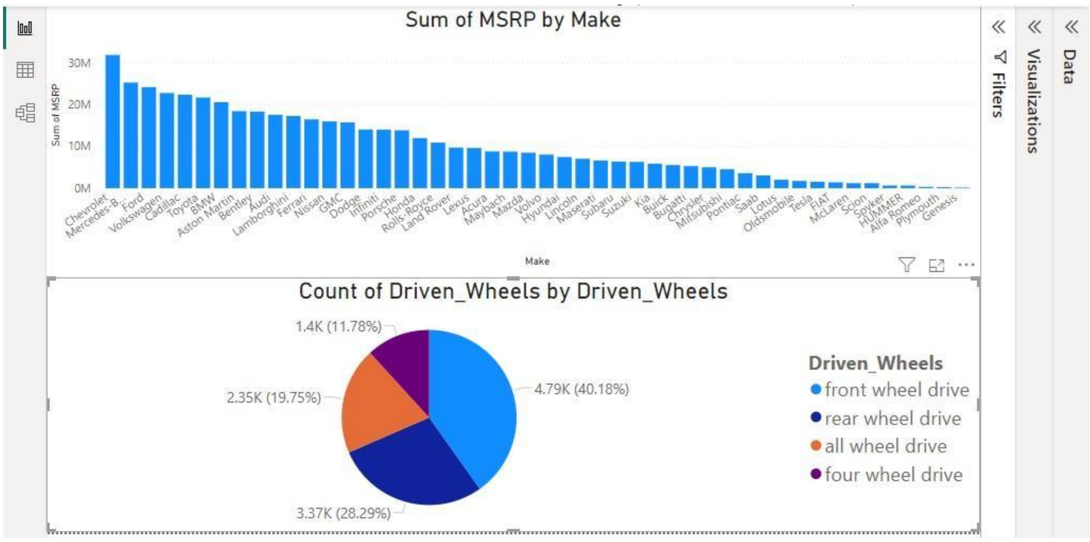
Set up automated data collection from the car prediction model and create a KPI dashboard using tools like Power BI.

Share the dashboard with the team and stakeholders to monitor progress and refine the model iteratively.

Iterate and Refine:
Continuously review the KPIs to ensure they align with evolving business objectives and improve the accuracy of car price predictions.

The car prediction KPIs are designed to align with the project's objective of accurately predicting car prices, enabling informed pricing decisions and inventory optimization.

Dashboard:


Sum of MSRP by Make
Sum of MSRP
30M
20M
10M
0M
Chevrolet
Mercedes-B...
Ford
Volkswagen
Cadillac
Toyota
BMW
Aston Martin
Bentley
Audi
Lamborghini
Ferrari
Nissan
GMC
Dodge
Infiniti
Porsche
Honda
Rolls-Royce
Land Rover
Lexus
Acura
Maybach
Mazda
Volvo
Hyundai
Lincoln
Maserati
Subaru
Suzuki
Kia
Buick
Bugatti
Chrysler
Mitsubishi
Pontiac
Saab
Lotus
Oldsmobile
Tesla
FIAT
McLaren
Scion
Spyker
HUMMER
Alfa Romeo
Plymouth
Genesis
Make
Filters
Visualizations
Data
Count of Driven_Wheels by Driven_Wheels
1.4K (11.78%)
2.35K (19.75%)
4.79K (40.18%)
3.37K (28.29%)
Driven_Wheels
● front wheel drive
● rear wheel drive
● all wheel drive
● four wheel drive