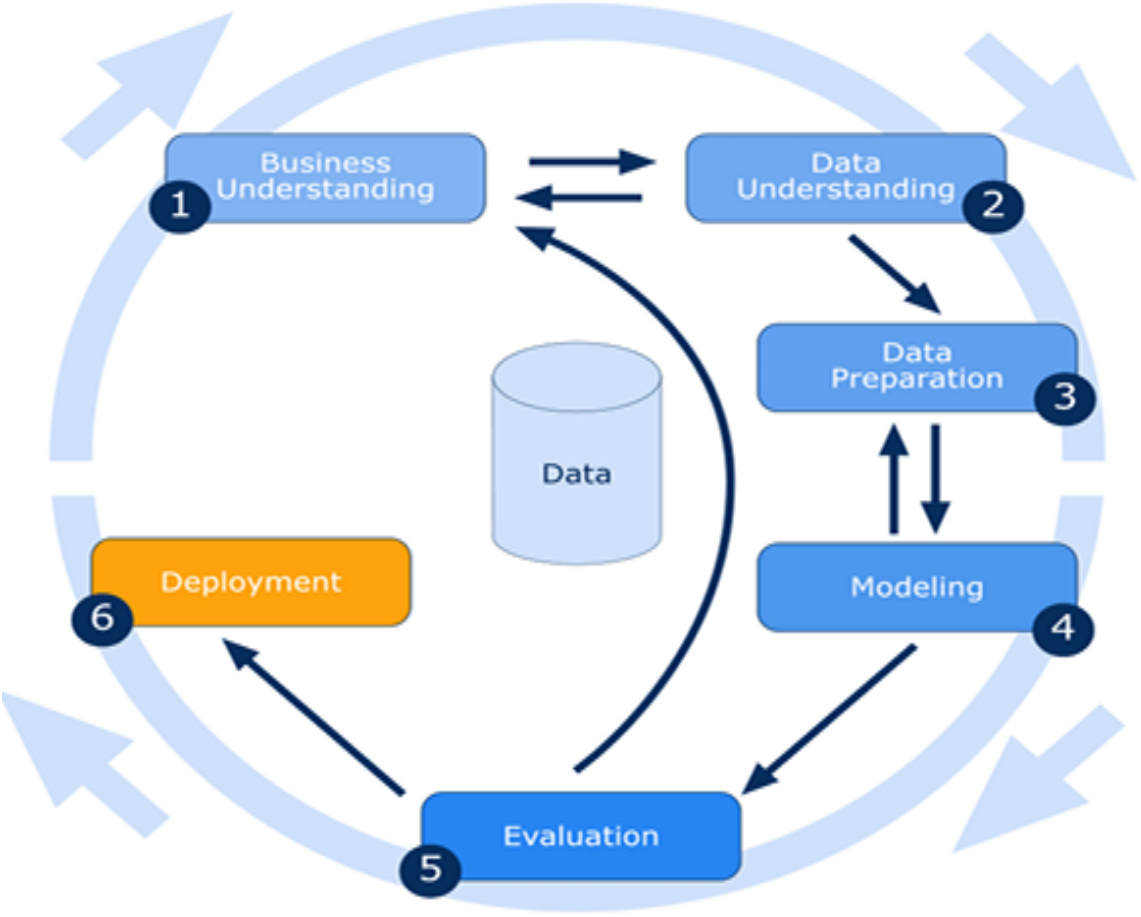
High-Level Design (HLD)

System Overview

The predictive system identifies vehicles likely to be 'bad buys' based on historical auction data, vehicle inspections, and other factors. The system assists buyers in avoiding costly purchases of 'lemon cars' without excluding too many good options.



Data Collection Module

- **Input:** CSV files from auctions, JSON from APIs.
- **Process:**
    - API integration with auction platforms.
    - Automated scrapers for historical auction data.
- **Output:** Consolidated dataset stored in a relational database (e.g., MySQL).

Data Processing and Cleaning Module

- **Functions:**
    - Handle missing values (mean/mode imputation).
    - Detect and remove duplicates.
    - Correct inconsistencies in formats (e.g., converting units like miles to kilometers).
- **Tools:** Python libraries like Pandas and NumPy.

EDA Module

- **Input:** Cleaned dataset.
- **Process:**
    - Generate visualizations: Scatter plots (age vs. price), bar charts (frequency of defects).
    - Identify relationships: Correlation matrix for numerical features.

- **Tools:** Matplotlib, Seaborn.

Feature Engineering Module

- **Input:** Processed dataset.
- **Process:**
    - Create derived features like depreciation rate.
    - Normalize numerical data.
    - Encode categorical data (e.g., One-Hot Encoding).
- **Output:** Feature matrix for model training.
- **Tools:** Scikit-learn preprocessing utilities.

Model Training Module

- **Input**: Feature matrix and target variable.
- **Process**:
    - Train models using historical data.
    - Perform hyperparameter tuning (e.g., GridSearchCV).
    - Use k-fold cross-validation for robustness.
- **Output**: Trained model saved in serialized format (e.g., .pkl).
- **Tools**: Scikit-learn, XGBoost, TensorFlow (if neural networks are needed).

Evaluation Module

- **Input**: Validation dataset.
- **Process**:

    - Calculate metrics:
        - **Precision**: Correct identification of bad buys.
        - **Recall**: Coverage of actual bad buys.
        - **F1-Score**: Balance between precision and recall.
    - Perform cost analysis to quantify savings.
- **Output**: Performance report.

Insights and Reporting Module

- **Input**: Predictions and auction results.
- **Process**:
    - Present cost savings and performance metrics.
- **Output**: Business intelligence insights for decision-making.

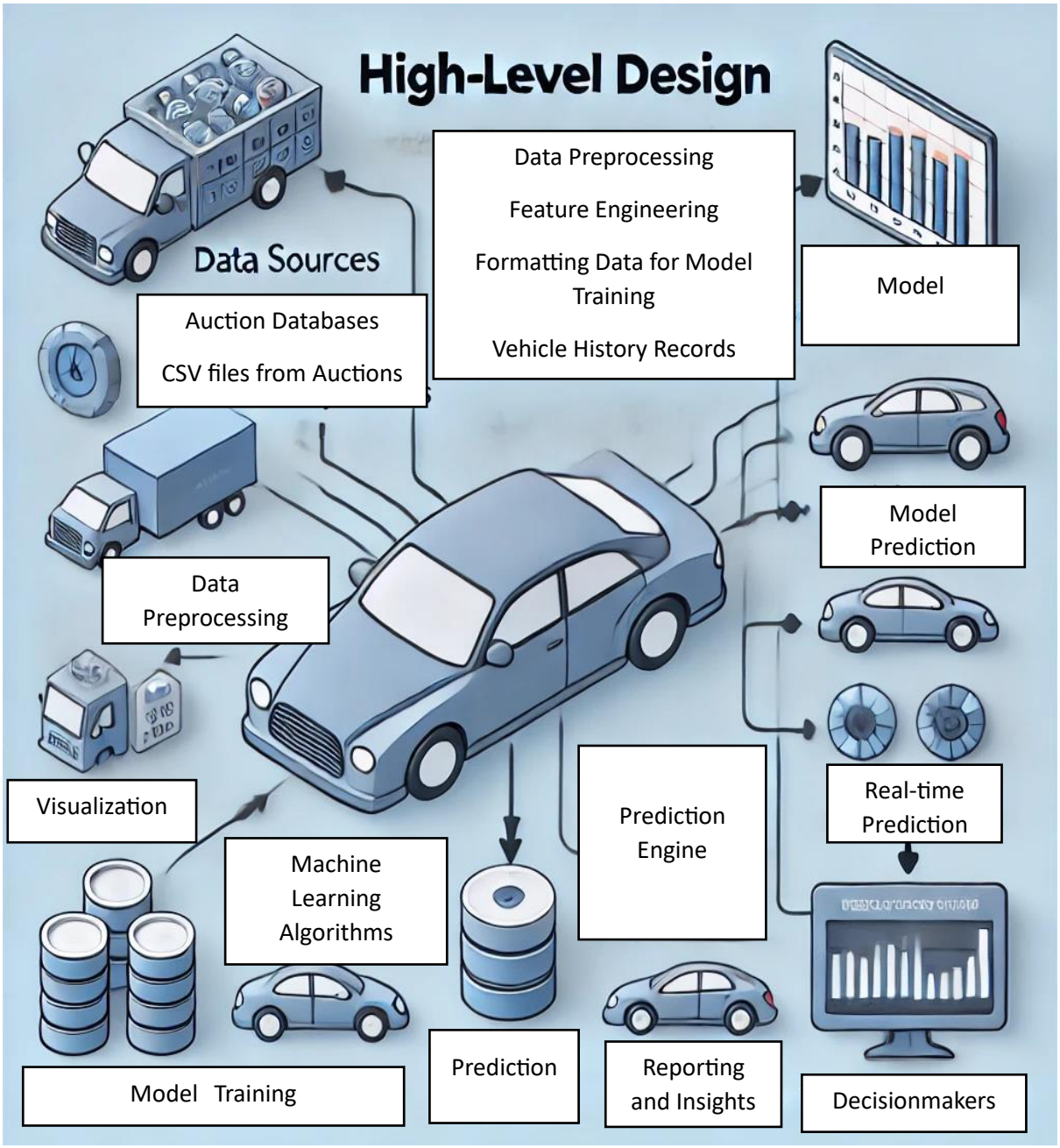


diagram represents the High-Level Design (HLD) for a predictive model identifying 'bad buys' in auto auctions. Here's a brief description of the components:

1. Data Sources: Includes auction databases, vehicle history records, and APIs providing vehicle data (age, mileage, condition, etc.).

2. Data Preprocessing: Handles cleaning, feature engineering, and formatting data for model training.
3. Model Training: Machine learning algorithms (e.g., Logistic Regression, Random Forest) are trained using historical data to predict 'bad buys.'
4. Prediction Engine: A real-time prediction system deployed on cloud platforms processes incoming data.
5. Reporting and Insights: Dashboards and metrics display predictions and key business insights for decision-making.