**Car24 (Car Rating Prediction)**

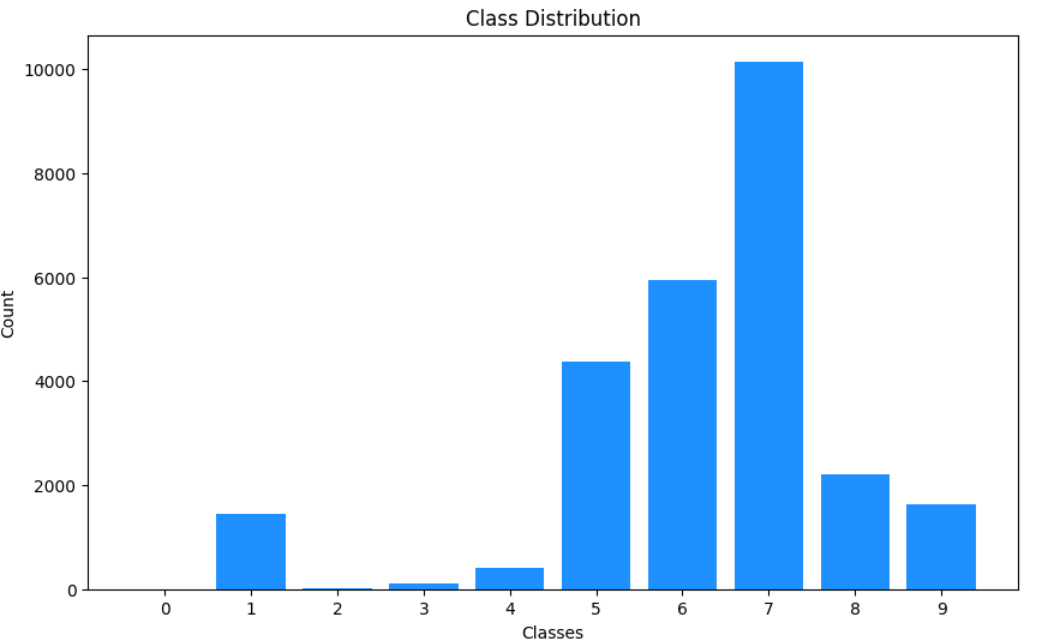
**Introduction:**

The task at hand is to develop a predictive model to estimate the engine rating based on inspection parameters provided in a dataset. This involves loading and preparing the data, conducting exploratory data analysis (EDA), building a predictive model, evaluating its performance, and conducting outlier analysis to identify cases where the rating may be incorrect. The goal is to showcase proficiency in Python/R programming, data science techniques, and the ability to deliver actionable insights within a commercial setting.

**Final best accuracy achieved : 83.3%**

**a. Steps Taken:**

**EDA:** Conduct exploratory data analysis to understand the distribution of variables, and visualize the rating distribution with in the dataset.



**rating engineTransmission**

**7 10152**

**6 5944**

**5 4379**

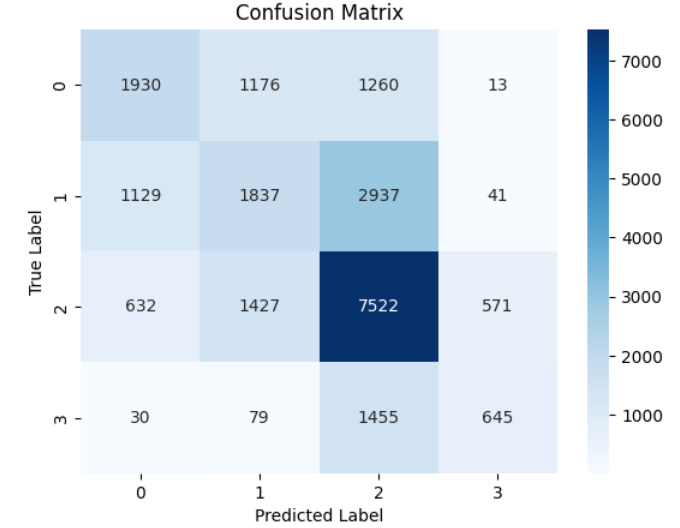
**8 2209**

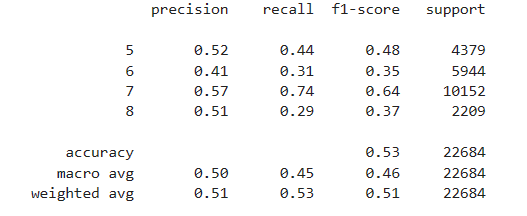
* Original data size: 22684
* Filtered data size (without mislabels): 11934
* Removed mislabeled cases: 10750

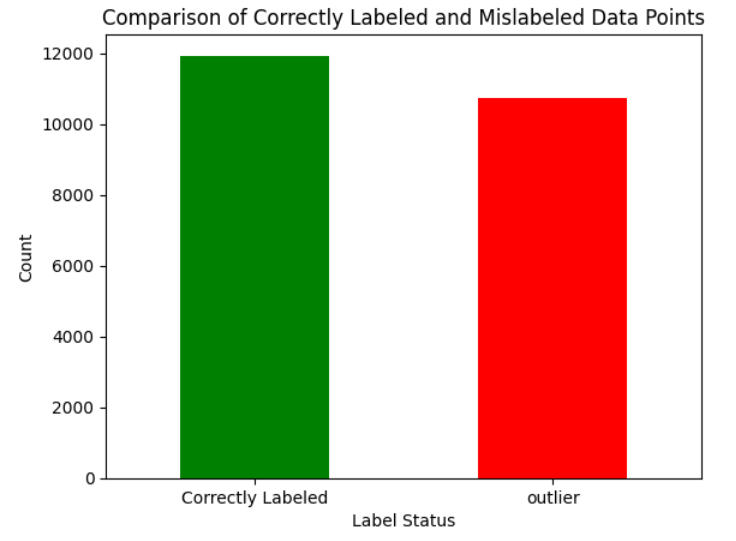
**Data Preparation Statistical Measures for Evaluation:**

* used **Chi-Square Testing** to check the dependency of categorical independent variable and target variable.
* Prepare the data for modeling, including handling missing values by replacing categorical variable with mode, and numerical with mean.
* perform label encoding for categorical variables, and splitting into training and testing sets.

**Model Building:** firstly tried Random forest classifier model (but it results in bad accuracy of around 53%) , and below is their confusion matrix and classification report.







* **after model prediction , I checked the outliers:**
* perform **PCA** to reduce dimensionality, and as data is unbalanced , I applied **Smote** to balance it,
* did hyperparameter tuning **(grid search)** and using **ExtraTreesClassifier** model for getting best accuracy of 83.3% .

**Encountered Issues:**

Challenges include handling missing data, selecting appropriate features, and addressing potential outliers in the dataset.

e. Insights Obtained:

Important features influencing engine rating are identified through feature importance analysis.

Visualizations such as scatter plots, and box plots help understand relationships between variables and detect patterns in the data.