**File Content for README.md**

**Car Price Prediction using Machine Learning**

**Objective**

The primary goal of this project was to build a robust machine learning model capable of predicting the selling price of a used car based on its features. This is a classic regression problem in data science. The project was completed as part of the OIB-SIP internship program.

**Dataset Overview**

The dataset, car data.csv, initially contained 301 entries with 9 features. An initial filtering step was performed to remove non-car vehicles, resulting in a clean dataset of 215 car entries for the analysis. After checking for and removing duplicate entries, the final dataset size was 213 rows.

Key Features:

The following features were used to predict the Selling\_Price:

* **Present\_Price**: The current showroom price of the car.
* **Driven\_kms**: Total kilometers the car has been driven.
* **Fuel\_Type**: Type of fuel (Petrol, Diesel, or CNG).
* **Selling\_type**: Whether the car is sold by a Dealer or an Individual.
* **Transmission**: Manual or Automatic.
* **Owner**: Number of previous owners.
* **Year**: The manufacturing year of the car.
* **Car\_Name**: The full name of the car model.

**Methodology & Steps Performed**

1. **Data Preprocessing & Feature Engineering:**
   * **Data Cleaning**: Non-car entries (such as bikes and scooters) were identified and removed, and duplicate rows were dropped.
   * **Feature Engineering**: A new feature, No\_of\_Years, was created from the Year column to represent the car's age, which is a key factor in price prediction. The Brand was also extracted from the Car\_Name column to simplify categorical analysis.
   * **Encoding**: Categorical features such as Fuel\_Type, Selling\_type, Transmission, Owner, and the newly created Brand were converted into a numerical format suitable for the model using One-Hot Encoding.
2. **Exploratory Data Analysis (EDA):**
   * Univariate analysis revealed that the distributions of Selling\_Price, Present\_Price, and Driven\_kms were right-skewed with some outliers.
   * Bivariate analysis showed a strong positive correlation between Selling\_Price and Present\_Price. It also highlighted a negative correlation between the car's age and Driven\_kms with the selling price.
   * A correlation heatmap confirmed Present\_Price as the most significant predictor.
3. **Model Building:**
   * The cleaned and engineered dataset was split into an 80% training set and a 20% testing set.
   * A **Random Forest Regressor** model was chosen for its robustness and ability to handle non-linear relationships in the data.
   * The model was trained on the training data (X\_train and y\_train) to learn the patterns between the features and the selling price.

**Model Performance and Outcome**

The model's performance was evaluated on the unseen test set, yielding the following results:

* **Mean Absolute Error (MAE):** 0.76 Lakhs. This means, on average, the model's predictions are off by ₹76,000 from the actual selling price.
* **Root Mean Squared Error (RMSE):** 1.47 Lakhs. This metric penalizes larger prediction errors more heavily.
* **R-squared (R2) Score:** 0.69. This indicates that the model's features explain approximately 69% of the variability in car selling prices, which is a good baseline for a predictive model.

Feature Importance:

The Random Forest model identified the most influential features for predicting car prices:

* **Present\_Price**: The current showroom price was the most important feature.
* **No\_of\_Years**: The car's age was the second most important factor.
* **Driven\_kms**: The total distance driven was also a key predictor.

**Tools & Libraries**

* **Python**
* **Pandas**
* **NumPy**
* **Matplotlib**
* **Seaborn**
* **Scikit-learn**

**The key insights from the project were clear:**

* A car's current price is the strongest indicator of its selling price.
* Car age and mileage are the top factors for price depreciation.
* Specific brands and features like fuel type significantly influence resale value.