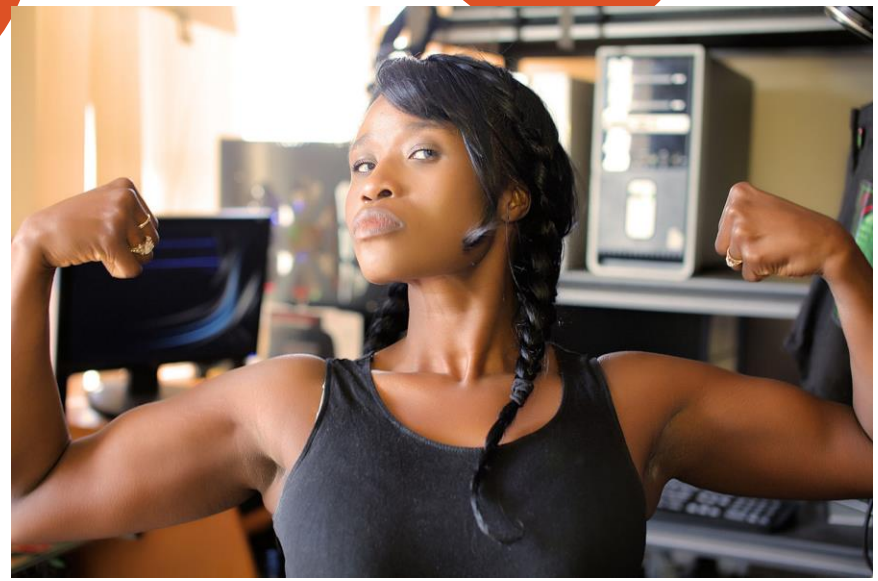
SNS Brownbook

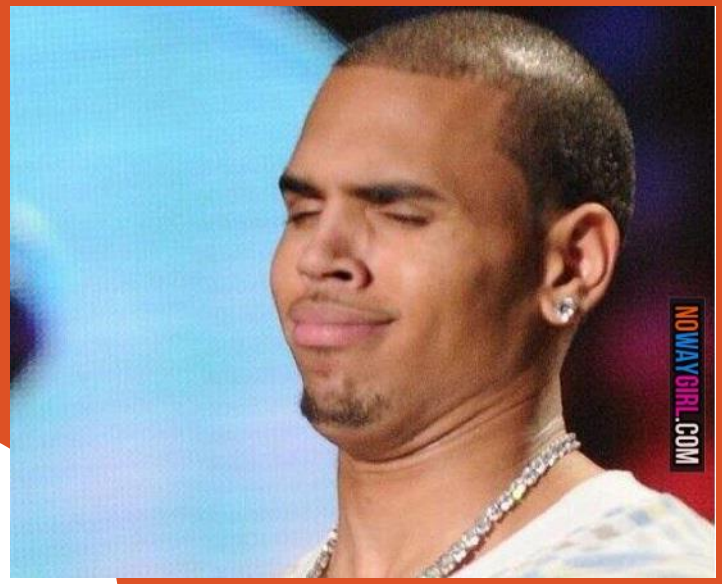
Used car solutions for your satisfaction

MEET THE TEAM



Sufi Hossain

Nyasha Gwaza



Saumili Chakravarty

Project Description

Problem:

Buying and selling used cars is often plagued by uncertainty and price manipulation by both fraudsters and dealerships. Our project uses machine learning to assess used car value based on attributes like mileage, brand, fuel efficiency, and engine specs. This enables sellers to make informed decisions (e.g., repair vs sell as-is), and gives buyers fair market price insight.

Solution:

Predict used car prices using supervised learning models

Compare traditional ML (Linear Regression, Random Forest) with a Neural Network

Evaluate models based on RMSE, R^2, and visual prediction accuracy

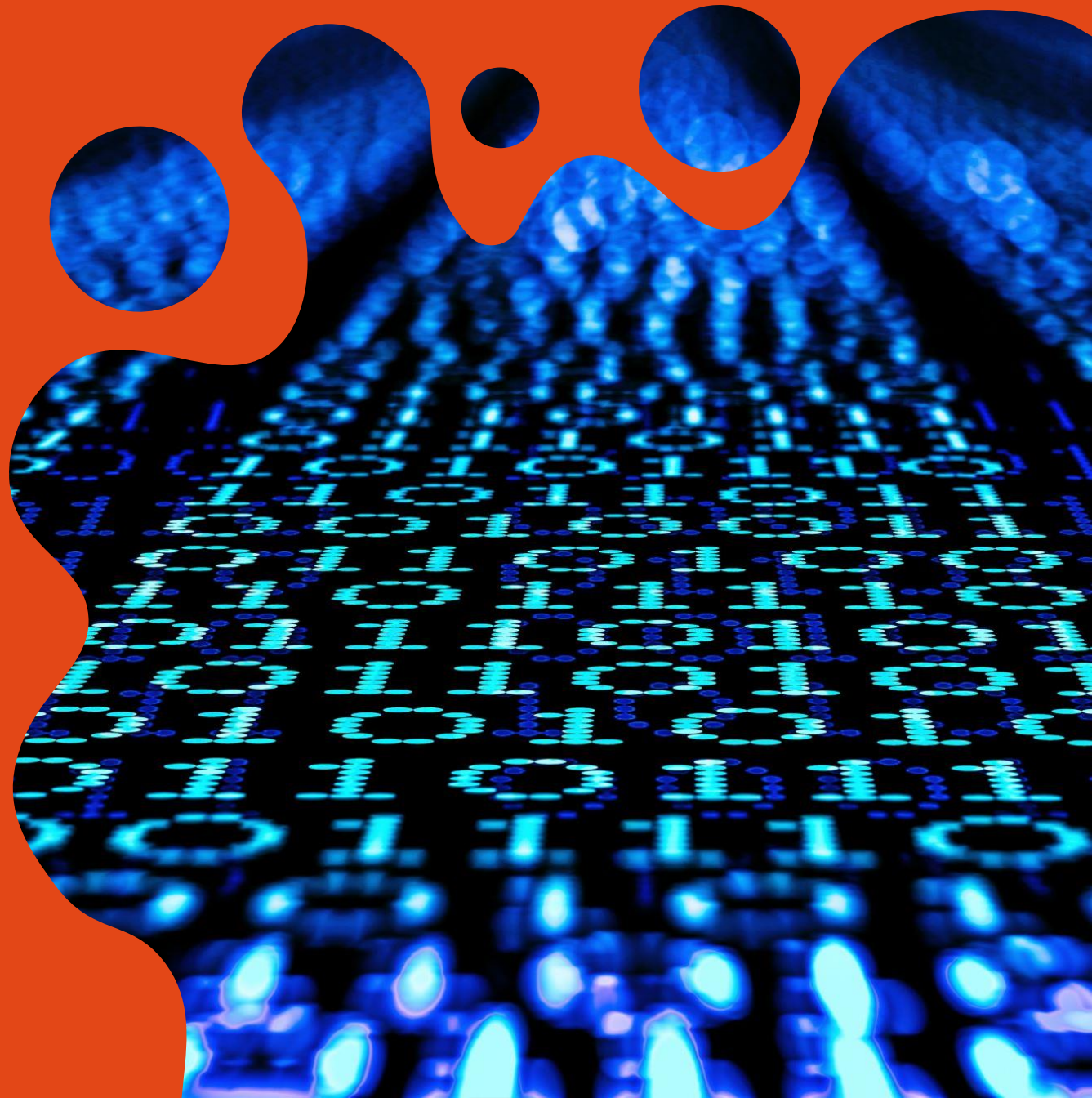
Data Collection

Data was collected from various U.S. zip using an instant Data Scraper tool, Auto Tempest.

We collected data from:

- Cars.com
- AutoTrader

Total data entries: 1138



Types of data

We collected the data based on these parameters.

- Brand, Car Model, Year
- Drive Train, Fuel Type, Engine Size
- Mileage, Battery (for EVs), City/Highway MPG
- Dealer Location (used to extract U.S. state)
- Target: Market price in USD

Data Cleaning and Preprocessing

```
print(data.dtypes)
```

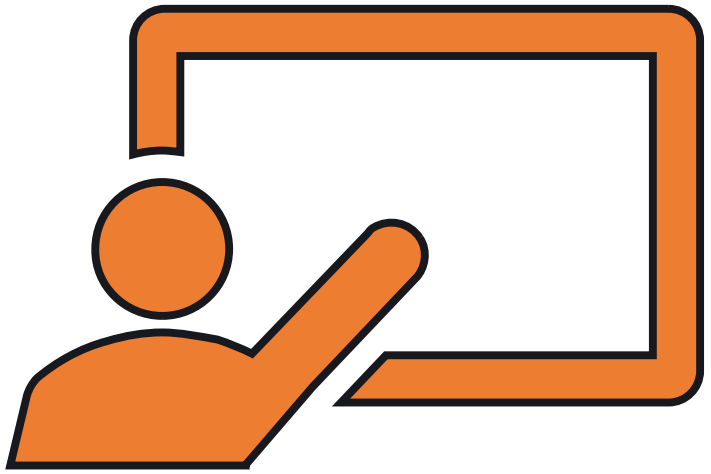
```
Year                    int64
Brand                  object
Drive Train            object
Fuel Type              object
Engine Type (L)       float64
Engine Type (Cyl)       int64
Battery (kWh)           int64
Mileage                 int64
Fuel (City              int64
HWY)                    int64
Price ($)               int64
dtype: object
```

Data Cleaning

- Dropped irrelevant columns such as car model, trim, reference link,
- Ensured that numerical features were correctly identified and read.
- Removed rows with missing values in any columns
- Had 899 data points after cleaning

Data Preprocessing

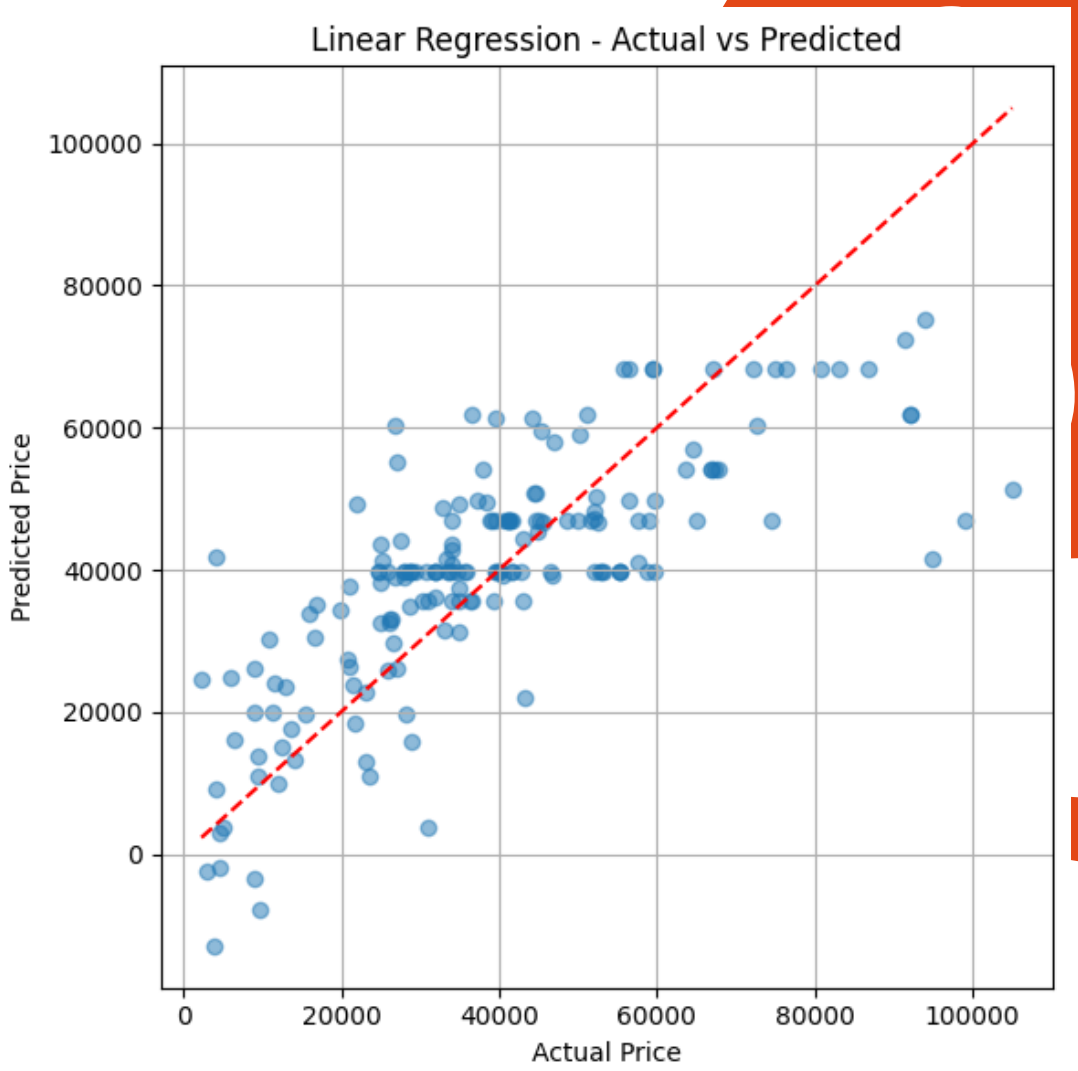
- Shuffled and split data using ratio 6:2:2
- Separated target Price from other features
- StandardScaler and OneHotEncoder scaling.
- Applied the same preprocessing pipeline to all datasets
- Converted outputs to dense arrays for modeling.



Training Models

Model Training Summary

Model	Key Info	RMSE & R^2
Linear Regression	Baseline model, assumes linear relationships	$13948.88 & 0.568
Random Forest	100 trees, nonlinear model, better fit	$13593.79 & 0.589
Improved Random Forest	3 dense layers, Adam optimizer, early stopping	$13685.90 & 0.584



Actual vs Predicted Car Prices (Linear Regression)

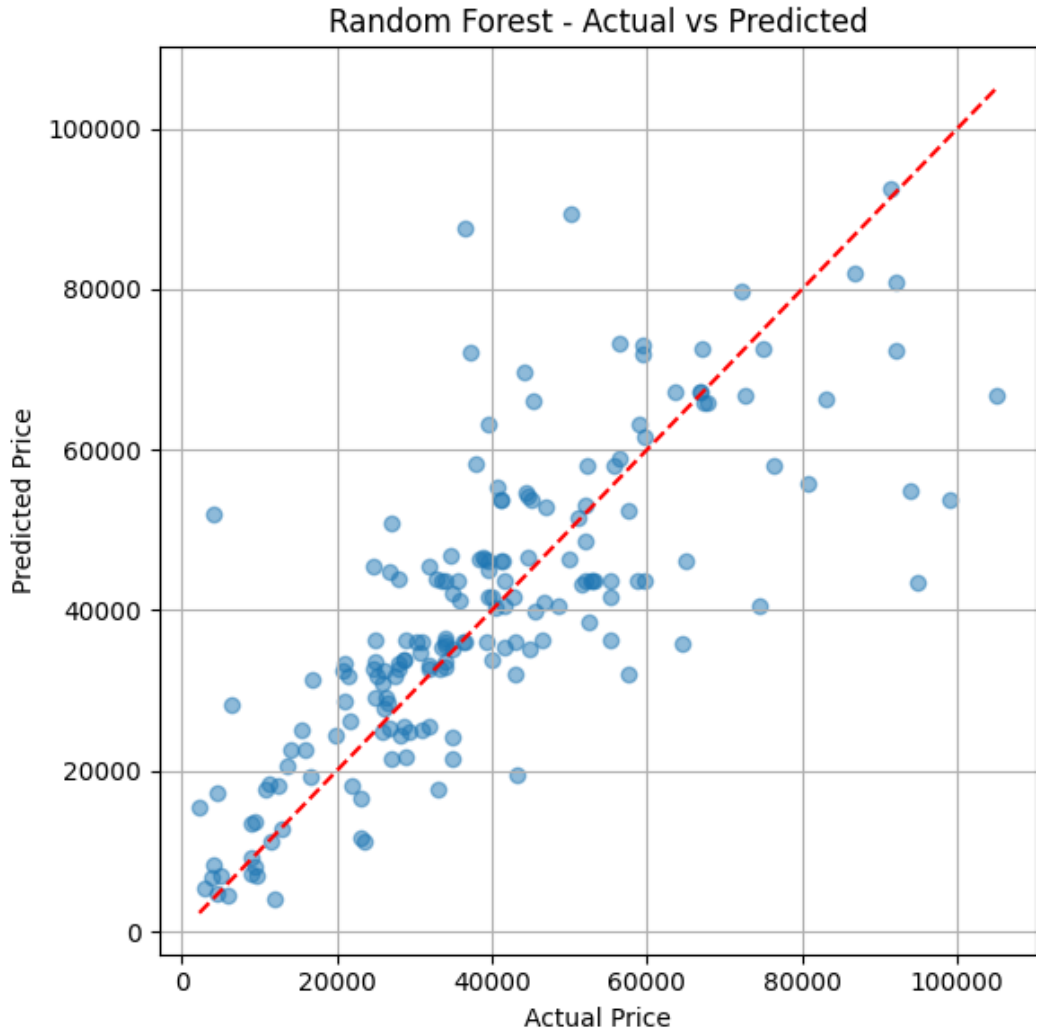
This plot is clear, but it shows that the model is badly underfitting.

Most of the predictions are very low compared to the actual prices.

The model is struggling to predict higher car prices.

Actual vs Predicted Car Prices (Best Random Forest)

- Captures nonlinear relationships
- Some overfitting on high-priced cars
- Performance slightly below linear regression



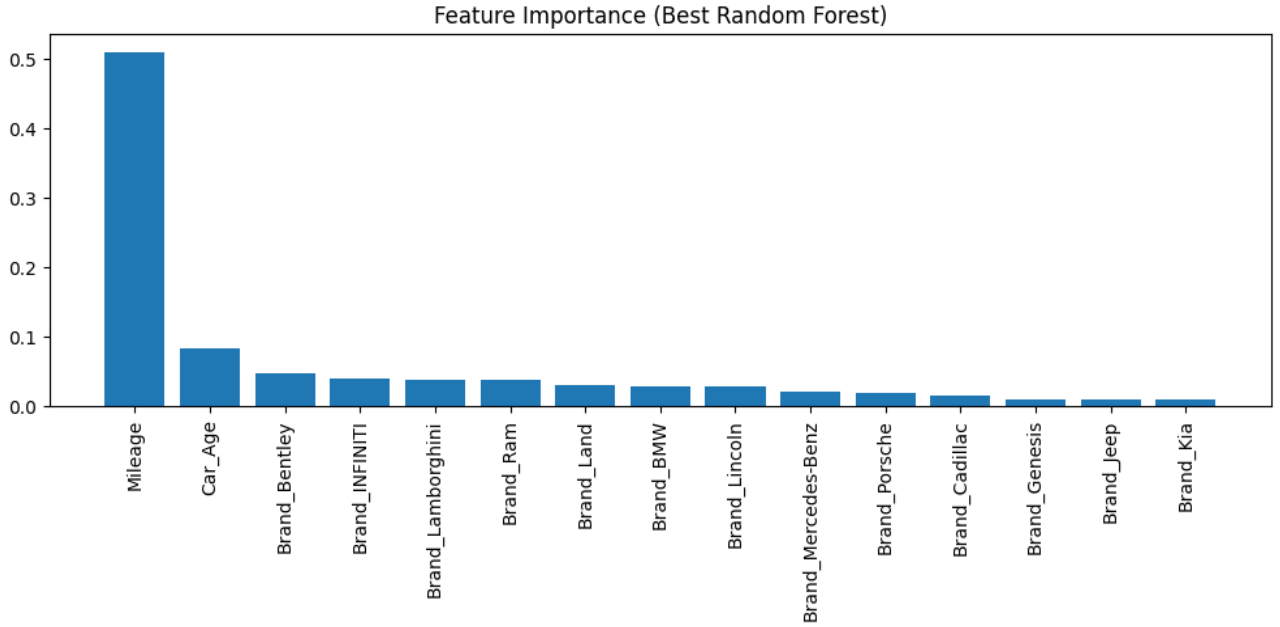
Results

Impacting Parameters

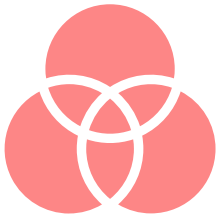
Mileage is by far the most important factor affecting used car price – consistent with real-world depreciation trends.

Car Age (derived from Year) was the second strongest predictor, reinforcing that newer cars generally retain higher value.

Brand and Drive Train also influenced predictions, though less significantly.



Summary



Best Model: Linear Regression (lowest RMSE and highest R^2 in this dataset)

Neural Network showed learning progress but was slightly outperformed

Data quality and feature variety are key to improving **performance**

Thank you!