Cars365

Used Cars Price Prediction using Spark

Group 3
December 8, 2022
Drishti Jain, Nisarg Patel, Nishi Morbia, Priyank Shah

Table of Contents

Cars are Depreciating Asset

How do you determine its resale value?

Dataset Description

- The dataset was obtained from **Kaggle**

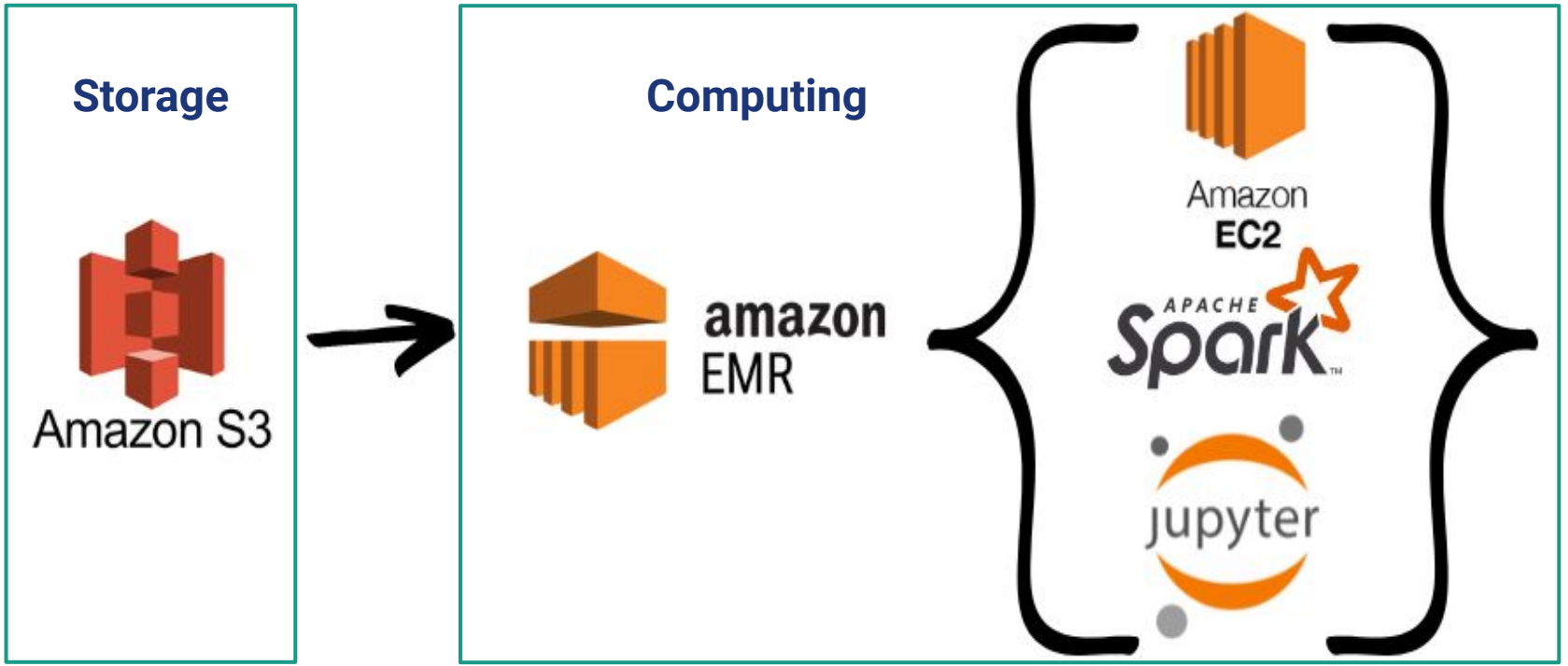
SIZE	RECORDS	COLUMNS
1.85 GB	*~3.2M*	*65*

- Columns are model of the car, brand of the car, city name, price, date listed, franchise dealer, kilometer driver and others

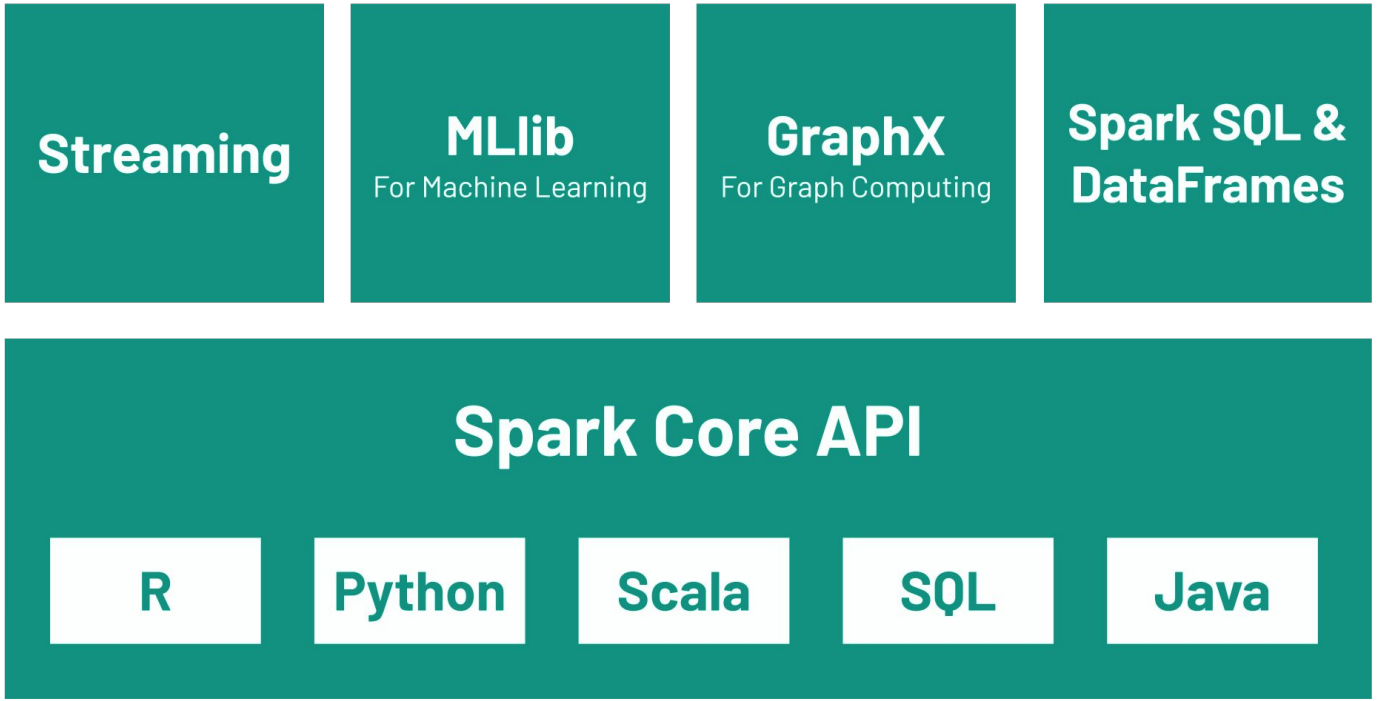
Use case and Business Questions

- Does the year of the car listed tell us anything about the used car **demographic**?
- Do we see a wide **distribution** of car prices in the market?
- How can a car dealer leverage horsepower enthusiasts?
- Do users care about **transmission type**?
- How **higher mileage affect the price** of the car?

Architecture



Why Apache Spark ?



Data Processing

Data Cleaning

- Filling in missing values/NaN values
- Fixing the values and data types of the columns
- Feature Engineering (creating new features from existing features)

Filling in missing values/NaN values

Percentage of missing values:

```
amount_missing_df = data.select([(count(when(isnan(c) | col(c).isNull(), c))*100/count(lit(1))).alias(c) for c in data.columns])
amount_missing_df.show()
```

```
+---+---+----------------+----------------+----------------+----------------+------------------+----------------+----+------------------+
|_c0|vin|     back_legroom|             bed|      bed_height|      bed_length|         body_type|           cabin|city| city_fuel_economy|
+---+---+----------------+----------------+----------------+----------------+------------------+----------------+----+------------------+
|0.0|0.0|4.880952752977353|99.37093553417354|88.72371476160863|88.72371476160863|0.39937624805077515|98.10468157712992| 0.0|16.358488620276937|
+---+---+----------------+----------------+----------------+----------------+------------------+----------------+----+------------------+
```

Percentage of missing values after cleaning:

```
amount_missing_df = data.select([(count(when(isnan(c) | col(c).isNull(), c))*100/count(lit(1))).alias(c)
for c in data.columns])
amount_missing_df.show()
```

```
+---+---+---------+----+------------------+------------+----------+----------+--------------+----------------+
|_c0|vin|body_type|city|city_fuel_economy|daysonmarket|dealer_zip|engine_type|exterior_color|franchise_dealer|
+---+---+---------+----+------------------+------------+----------+----------+--------------+----------------+
|0.0|0.0|      0.0| 0.0|               0.0|         0.0|       0.0|       0.0|           0.0|             0.0|
+---+---+---------+----+------------------+------------+----------+----------+--------------+----------------+
```

Fixing the values and data types of the columns

Converting boolean values to numeric values

```
from pyspark.sql.types import IntegerType,BooleanType,DateType
import pyspark.sql.functions as F

def fromBooleanToInt(s):
    if s == True:
        return 1
    elif s==False:
        return 0
    else:
        return None

fromBooleanToInt_udf = F.udf(lambda x: fromBooleanToInt(x), IntegerType())

data = data.withColumn('fleet',fromBooleanToInt_udf(data['fleet']))
data = data.withColumn('isCab',fromBooleanToInt_udf(data['isCab']))
data = data.withColumn('is_cpo',fromBooleanToInt_udf(data['is_cpo']))
data = data.withColumn('is_new',fromBooleanToInt_udf(data['is_new']))
data = data.withColumn('is_oemcpo',fromBooleanToInt_udf(data['is_oemcpo']))
data = data.withColumn('salvage',fromBooleanToInt_udf(data['salvage']))
```

Fixing the values and data types of the columns

Checking value counts to look for number of null values

```
data.groupBy('length').count().orderBy('count', ascending=False).show()
```

```
+--------+-----+
|  length|count|
+--------+-----+
|    null| 2497|
|189.8 in| 1419|
|191.7 in| 1271|
|183.1 in| 1073|
|184.5 in| 1001|
|176.4 in|  989|
|231.9 in|  933|
|182.3 in|  818|
|180.5 in|  786|
|192.5 in|  679|
|188.8 in|  613|
|182.1 in|  599|
|198.8 in|  572|
|183.5 in|  499|
```

Converting string to numeric values, filling null values with mean of the column

```
data = data.replace({'--':None}, subset=['wheelbase'])
data = data.withColumn("wheelbase",split(data["wheelbase"]," ").getItem(0))
data.withColumn("wheelbase",col("wheelbase").cast("float"))
df_mean = data.select(mean(col('wheelbase')).alias('avg')).collect()
avg = df_mean[0]['avg']
data = data.withColumn('wheelbase', when(data.wheelbase.isNull(),
                                         lit(avg)).otherwise(data.wheelbase))
print(data.filter(col("wheelbase").isNull()).count())
```

```
0
```

Feature Engineering

```
# Combining front_legroom and back_legroom
data = data.withColumn('legroom', expr("front_legroom + back_legroom"))
```

```
data = data.withColumn("major_options_count",size(split(col("major_options"),",")))
data.groupBy('major_options_count').count().orderBy('count', ascending=False).show()
```

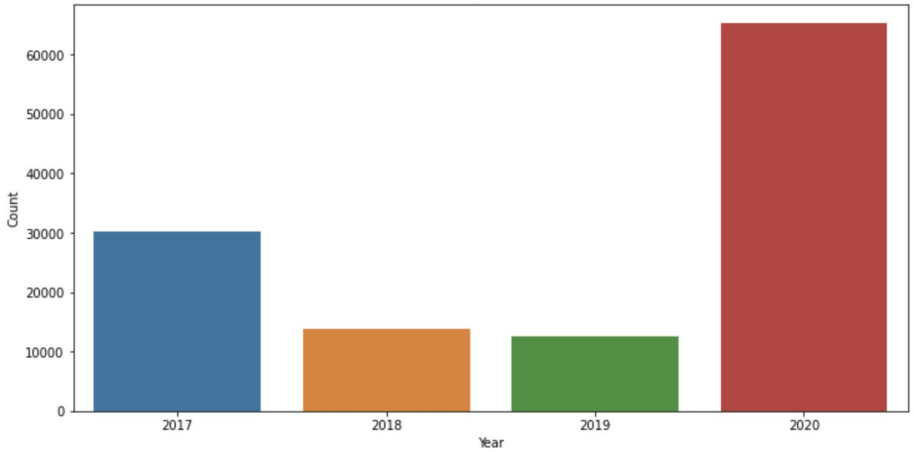
```
+--------------------+-----+
|       major_options|count|
+--------------------+-----+
|                null| 3983|
|['Alloy Wheels', ...| 1331|
|    ['Steel Wheels']|  826|
|    ['Alloy Wheels']|  799|
|['Bluetooth', 'Ba...|  732|
|['Steel Wheels', ...|  548|
|   ['Backup Camera']|  519|
+--------------------+-----+
```

```
+-------------------+-----+
|major_options_count|count|
+-------------------+-----+
|                  4| 5415|
|                  5| 4752|
|                  3| 4578|
|                  6| 4540|
|                  7| 4094|
+-------------------+-----+
```

Business findings with Exploratory Data Analysis (EDA)

Cars per year

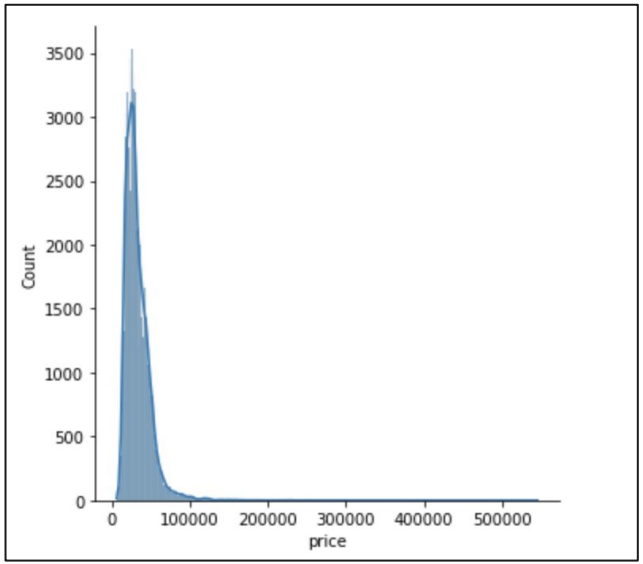
- **2020** marked a significant increase in the sales of used cars over the US
- From 2017 to 2020 the number of used cars for sale in market **increased by 133%**



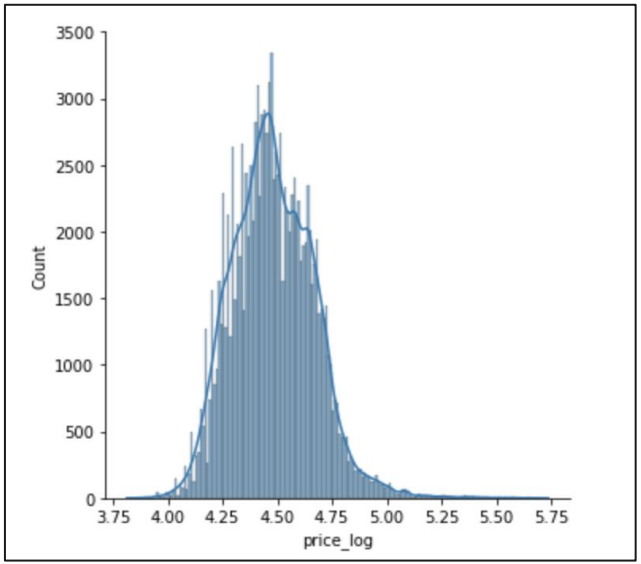
Cars per year

Price Distribution

- For our dataset the price variable - that is supposed to be our target variable was heavily skewed to right.
- Majority of the cars were in the range of **$ 2000 to $ 150,000,** with outliers ranging from $ 200,000 to $ 500,000

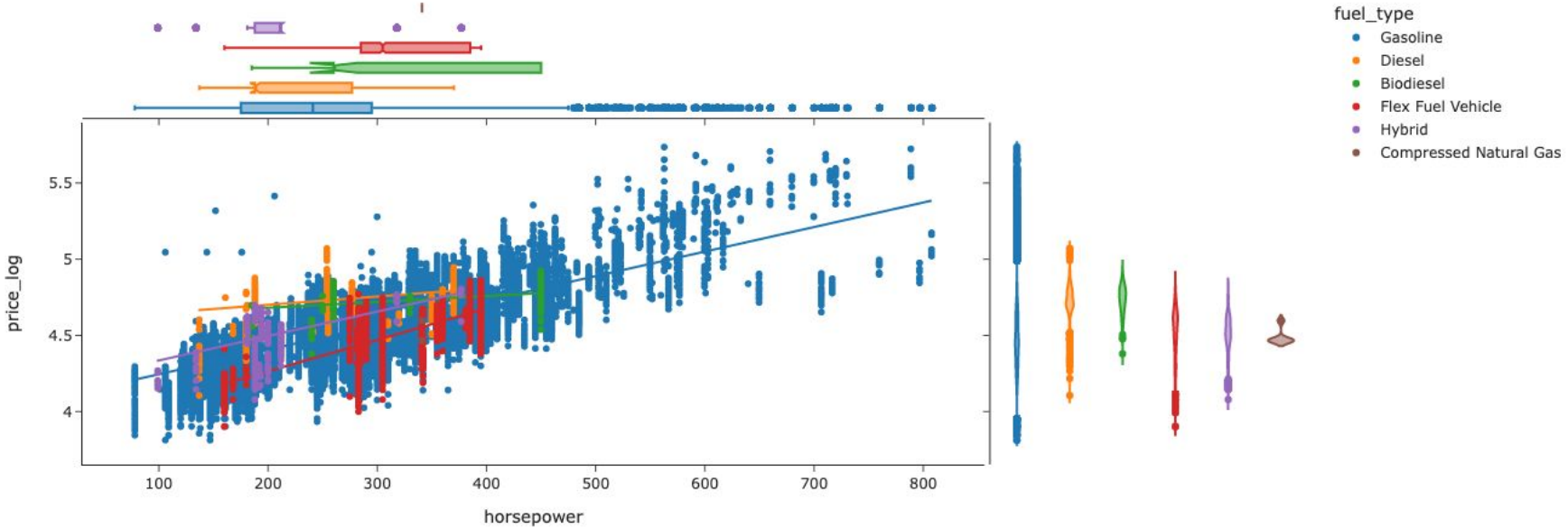


Log Transform



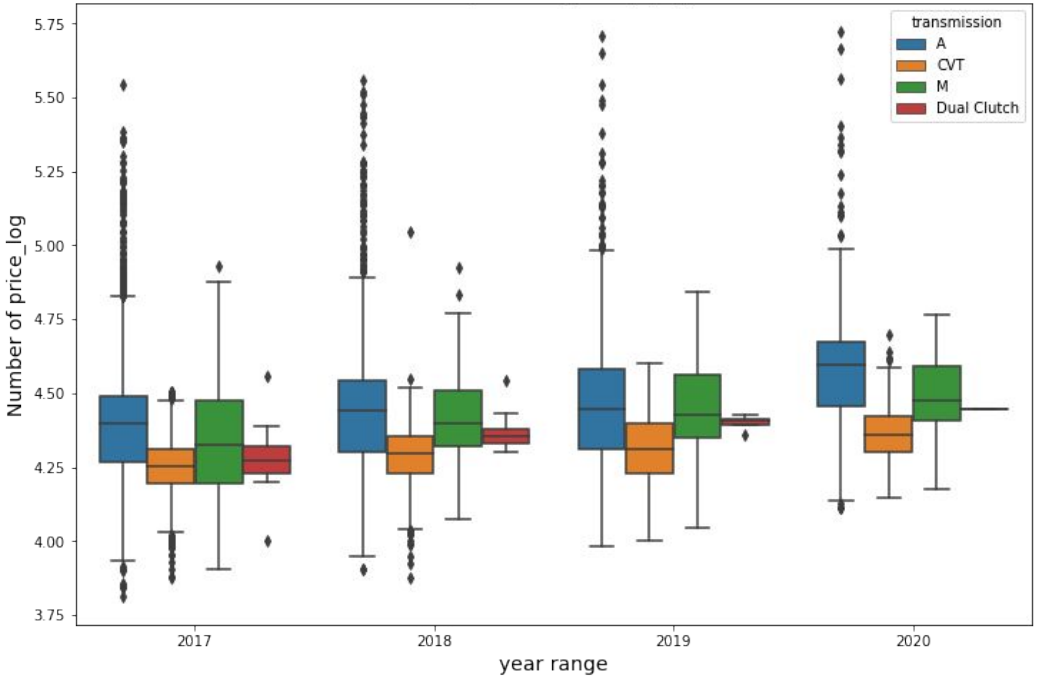
Price vs Fuel type vs Horsepower

- Gasoline & Horsepower enthusiasts are in for a treat with wide distribution of cars.
- We also conclude from the graph below that, for all fuel types, with increase in horsepower, the price of car also increases.



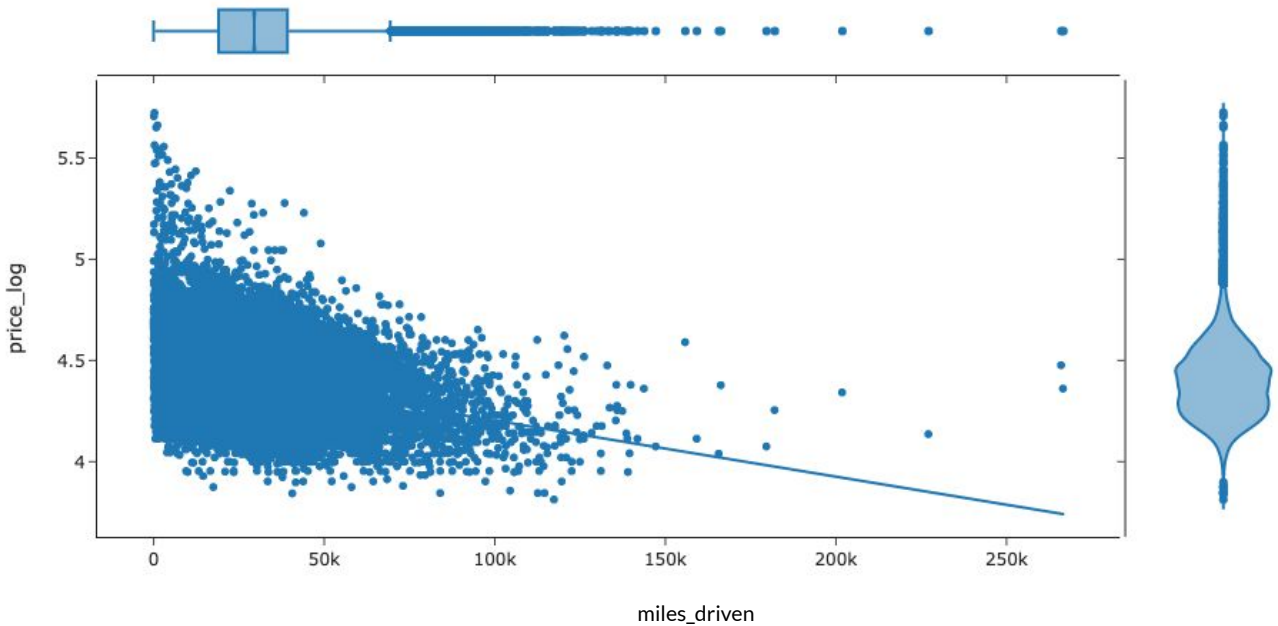
Price vs Transmission

- Automatic vehicles are priced more competitively, followed by manual ones.
- Year-to-year listed price also increases for all the transmission types.



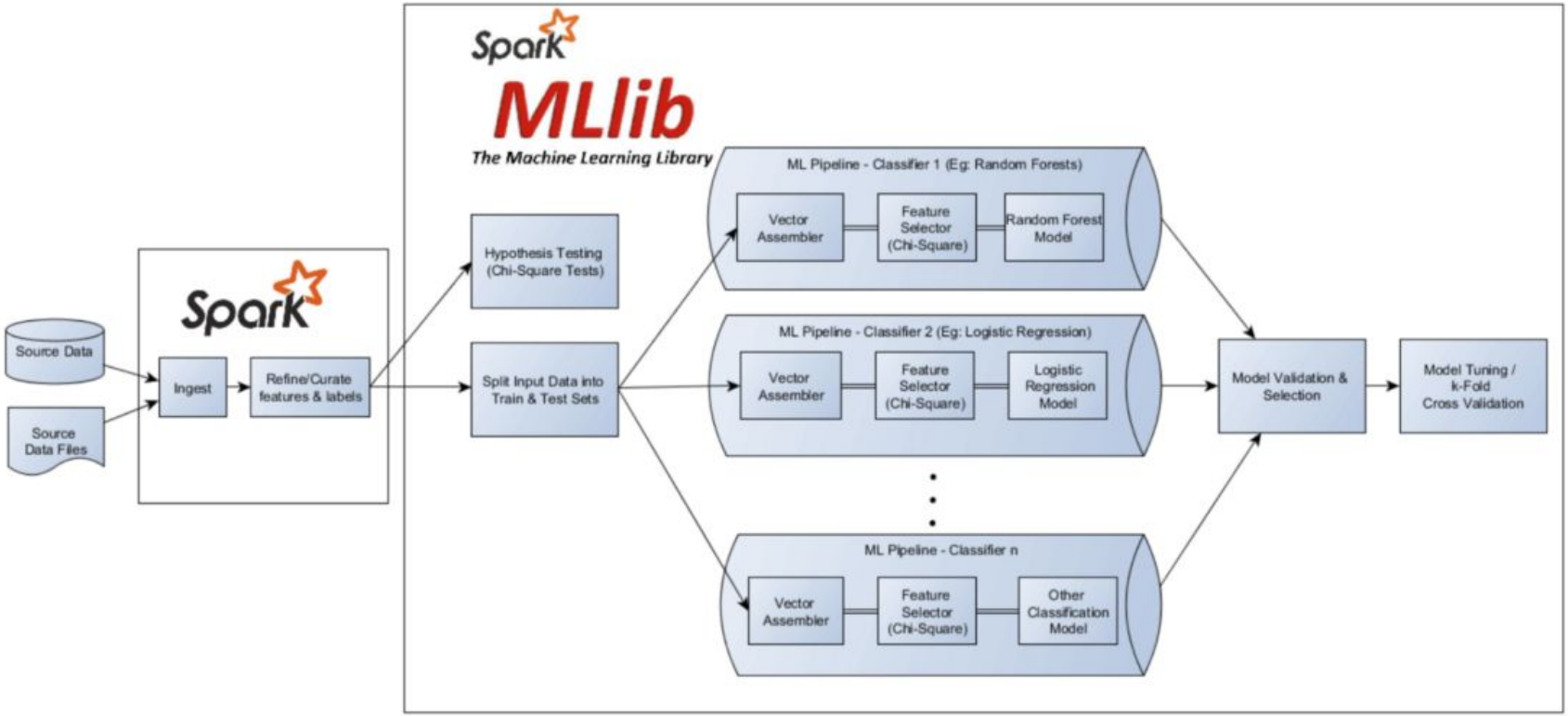
Price vs Miles Driven

- Evidently, as the used car's mileage travelled increases, the price of car drops.
-



Data Modeling

Regression pipeline



Modeling Process

1

Feature Engineering

Creating new features by transforming existing features including binning, dummy variables and other transformations

2

Split into Train-Test

Split the dataset into train and test in 70:30 proportion

3

Vector Assembler

VectorAssembler is a transformer that combines a given list of columns into a single vector column.

4

Train the model

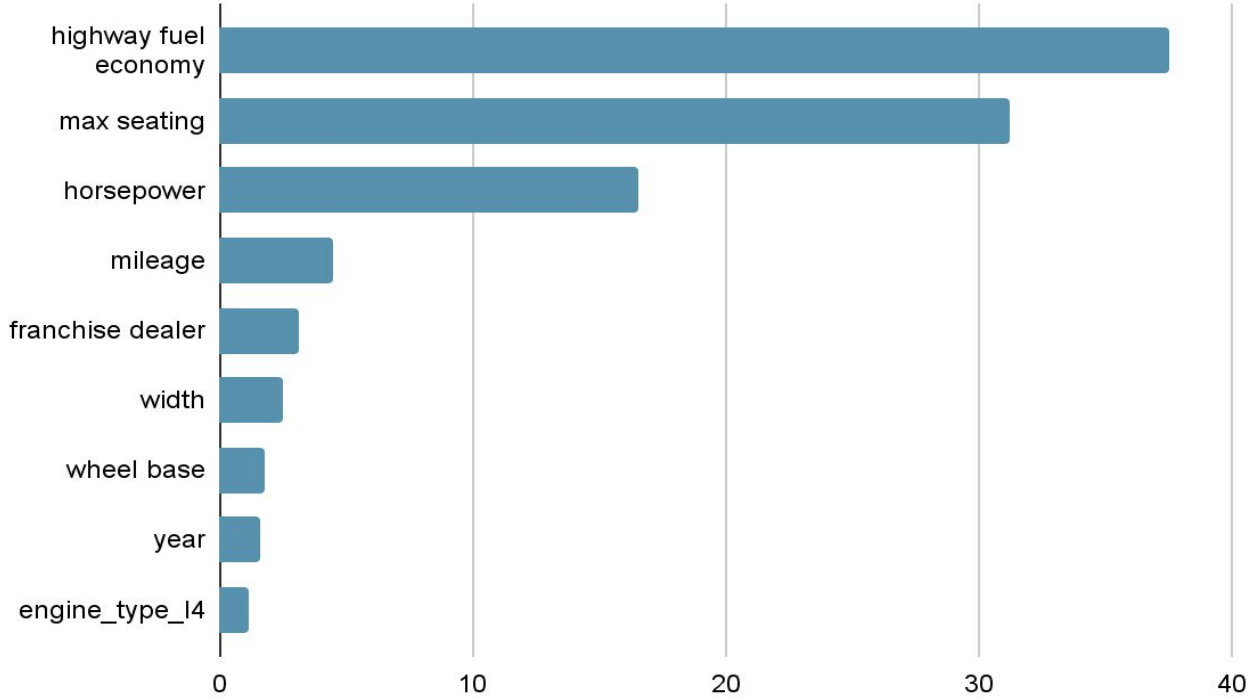
Using pyspark machine learning libraries, train various regression models

5

Evaluate the model

Compare performance (R squared, RMSE) on train and test data for all models

Feature Importances



Model Performance

Model Name	Train data RMSE	Test Data RMSE	R-square value
Linear Regression	11,798	13,458	69%
Decision Tree Regression	13,056	15,191	64%
Gradient Boosting Regression	11,006	13,036	71%

Putting it all together

Putting it all together

We can conclude that a car dealer can use our model built using **Gradient Boost** to determine the price of the used cars.

Does the year of the car listed tell us anything about the used car demographic?

- Yes, earlier we saw that as the years progresses, more number of used cars are listed in the market. Specifically the 2020's Covid era, show a significant increase in the number of cars

Do we see a wide distribution of car prices in the market?

- No. Most of the cars on the market have prices that attract a new buyer/ middle class families, as per the price range listed earlier, though there are some outliers.

Putting it all together

Do users care about fuel type?

- Yes, from fuel-type vs price we can derive that the market is dominated by Gasoline cars, supporting the demand/ supply characteristics

How can a car dealer leverage horsepower enthusiasts?

- Increase in horsepower marks higher priced vehicles. So the car dealer can effectively match price with the horsepower of the car with respect to the market

How higher mileage affect the price of the car?

- It is concluded that, buyers will pay more to get a lesser driven car. For a car dealer it is recommended to keep the cars that are driven below 40k miles for him to sell it with a good price.

Thank You!

Fun Game:
bit.ly/pokemonorbigdata