KLYM

**Data Science Director**
Technical home test and business assessment

Candidate: Efraín Galvis


**Agenda:**

1. Business Assessment
2. Data Analysis & Model Building


# Business Assessment

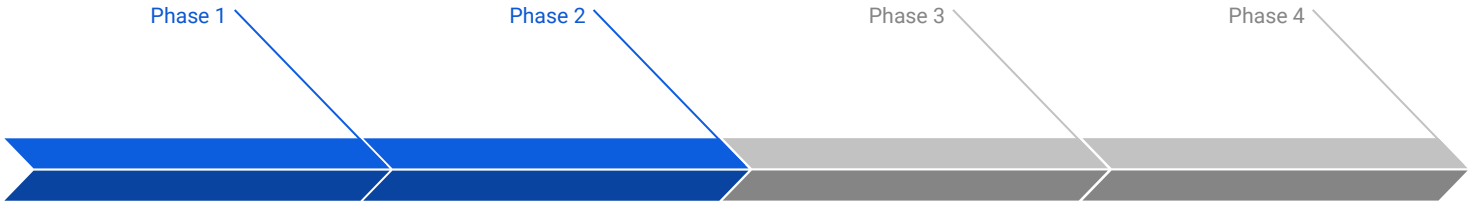
# Business Assessment

**Goal:** Demonstrate creativity and problem-solving skills in addressing the given problem.

**Business Case:** The Innovation and Sales departments of a car dealership in the USA are working on modernizing their operations by developing a model for automatic car pricing solutions for their customers.

KLYM

# Methodology




**Exploring Data**

Understanding the problem from a data perspectiv

**Hypothesis**

Making questions to the data

**Transforming fields**

Putting all in place so the models can understand well the patterns and insights

**Modeling, Tuning, and Iterating**

The chosen models were Random Forest and XGBoost.

# Results and Insights

## What determines the price of a vehicle?

- Odometer
- Year
- Manufacturer
- Condition
- Cylinders
- Fuel
- Transmission
- Drive
- Type of vehicle
- Paint color
- State

| Model | MAE | RMSE | R-squared |
|---|---|---|---|
| Random Forest | 4352.87 | 7225.47 | 0.58 |

**Understanding the metrics**

**MAE:** Indicates that, on average, the pricing predictions deviate from the actual market prices by approximately USD$ 4,352.87. This level ensuring that the cars are competitively priced in the market.

**RMSE :** This means that the majority of pricing predictions are within a range of approximately USD$ 7,225.47 of the actual market prices.

**R-squared:** Indicating that it explains 58% of the variance in car prices. This level of explanation demonstrates that our model captures crucial factors influencing pricing decisions.

The advantages of this model are that it allows us to create a specific margin for discounts and a range for negotiation.

KLYM

# During the process of finding the best model

**We explore different alternatives with other models and transformations**

| Iteration | Model | MAE | RMSE | R-squared |
|---|---|---|---|---|
| Hyperparameter tunning | Random Forest | 4353.11 | 7224.88 | 0.58 |
| | XGBoost | 4574.37 | 7366.42 | 0.56 |
| Scaling variables | Random Forest | 4352.87 | 7225.47 | 0.58 |
| | XGBoost | 4574.37 | 7366.42 | 0.56 |
| Frequency encoding | Random Forest | 4557.94 | 7509.52 | 0.55 |
| | XGBoost | 4582.42 | 7410.25 | 0.56 |
| Excluding Paint Color and State fields | Random Forest | 4536.05 | 7420.32 | 0.56 |
| | XGBoost | 4672.60 | 7563.29 | 0.54 |

KLYM

# Pricing Strategy

## Loss Scenario

Let's consider a scenario where our model predicts a car's price to be $30,000, but the actual market price is $32,000.

**MAE:** Our pricing predictions are off by approximately $4,352.87. In this scenario, the error is $2,000 below the actual price.

Our strategy acknowledges this and ensures the sale is finalized. While we incur a loss of $2,000 due to underpricing, we guarantee the sale and maintain customer satisfaction.

## Profit Scenario

Let's consider a scenario where our model predicts a car's price to be $40,000, but the actual market price is $38,000.

**MAE:** Our pricing predictions are off by approximately $4,352.87. In this scenario, the error is $2,000 above the actual price.

Our strategy takes advantage of this opportunity. We price the car at $2,000 above market value, aiming to maximize profit while remaining competitive.

# How likely are our customers to buy?

The probability that the customer will still buy the car is: $P(Buying|Price_{Real}) = 1 - min(1, \frac{Price_{Predicted}}{Price_{Real}} - 1)$

**Using our model:**

- P(Buying|PriceReal) = 1 - min(1, $44.352,87 / $40,000 - 1)
- min(1, 1,1088 - 1) = min(1, 0,1088)
- P(Buying|PriceReal) = 1 - 0,1088 = 0.8912

In the profit scenario where our model overestimates the car's price by $4.352,87, **there is a 89,12% probability that the customer will still buy the car.**

**Using the last example:**

- P(Buying|PriceReal) = 1 - min(1, $40,000 / $38,000 - 1)
- min(1, 1.0526 - 1) = min(1, 0.0526)
- P(Buying|PriceReal) = 1 - 0.0526 = 0.9474

In the profit scenario where our model overestimates the car's price by $2,000, **there is a 94.74% probability that the customer will still buy the car.**

KLYM


# Data Analysis & Model Building

# Data Analysis

With the variables Odometer, Condition and VIN. Address the following items

1. Evaluate the completeness of the variable and highlight any findings you think relevant for a discussion.
2. Evaluate how well the value discriminates your dependent variable and determine if you would use it in your model.
3. If the previous answer was affirmative then how would you propose to include it in your model. That is, what transformations, techniques, or tools would you use to include it.

Are there any duplicate, redundant or irrelevant variables in your data set? If there are, how would you deal with them?

KLYM

# Data Analysis

1. Evaluate the completeness of the variable and highlight any findings you think relevant for a discussion.

| Variable | Description |
|---|---|
| Odometer | 18,31% of the data is null. |
| | Outliers where identified and removed using Z-score |
| | Unrealistic records. Odometer readings below 50 miles (80 km) from years before 2008, were removed. |
| | After pruning outliers, the missing values were filled using the median |
| Condition | 43,12% of the data is null. |
| | Missing values were filled with "missing_unkwn" |
| VIN | 40,2% of the data is null. |
| | 5% of the data is duplicated |
| | 98% of the VINs have a length of 17 characters. This is a strong indicator that possible future VINs under 17 characters long should be reviewed. |
| | 20 unusual VINs were found, (e.g. Van Truck Center, IDRIVEFRESNO.COM) |

# Data Analysis

2. Evaluate how well the value discriminates your dependent variable and determine if you would use it in your model.

After eliminating outliers in the target variable, the correlation between Condition and Odometer reveals significant patterns.

| Variable | Description |
|---|---|
| Odometer | We observe a **weak negative correlation of -33%** between price and odometer, indicating that as the odometer reading increases, the price of the car tends to decrease. |
| Condition | After performing an ANOVA test, we identified a p-value of 2.682770381619786e-101, which is less than the significance level of 0.05. Therefore, we conclude that **the condition of the cars directly impacts the pricing of the vehicles.** |
| VIN | VINs should serve as unique identifiers for vehicles (id); however,the current state of this field is corrupted and **should not be used as an ID.** |

KLYM

# Data Analysis

3. If the previous answer was affirmative then how would you propose to include it in your model. That is, what transformations, techniques, or tools would you use to include it.

| Variable | Description |
|---|---|
| Odometer | Perform standardization (z-score) |
| Condition | one-hot encoding |

KLYM

# Data Analysis

Are there any duplicate, redundant or irrelevant variables in your data set? If there are, how would you deal with them?

- Geographical variables such as 'region,' 'latitude,' 'longitude,' and 'state' seem to be redundant. For simplicity purposes, we will only keep 'state' which only have 51 categories
- The fields 'fuel' and 'transmission' can easily imputed with the most frequent categories
- The 'drive' field, we can completed using the 'model' or 'description' fields.
- Missing values in the 'cylinders,' 'type,' and 'paint_color,' will be filled with 'not specified.'
- Values not included in the modeling:
    - 'size,' as 70% of the data contains missing values.
    - 'title status,' as 95% of the data falls into the 'clean' category.
    - 'description,' which currently provides no additional information except for populating the 'drive' field.
    - ‘vin,' as it does not provide useful information.
    - 'model' field is corrupted and may require a thorough cleansing process. Currently, it contains 6,298 different categories.

KLYM

# Model Building

Discuss how would you monitor your model in 3 months time to determine its validity. (No Code)

1. **Continuous Monitoring for Data Distribution Shifts**
2. **Automated Alert System for Performance Metrics**
3. **Leveraging User Feedback for Model Insights**

KLYM


Thank you

Efraín Galvis