# Oblig ML DAT158 HT2024

## Members: Leonard Heldal (669778), Severin Johannessen (669799), Iver Thoresen Malme(669820)

#Gruppe: 21

## Project Name: Car Price Prediction

## Introduction

The purpose of this project is to develop a car price estimation service that provides users with a reliable estimate of a car’s current market value, requiring no login or sharing of sensitive information. The target users are private individuals seeking insight into their car's market value and professional stakeholders, such as car dealers, who can benefit from better pricing and market adjustments.

## Business Objectives

The service is designed to meet several key business objectives:

1. * **Optimize Pricing**: Provide realistic car price estimates to dealers and sellers that reflect current market conditions, allowing them to set competitive prices.
2. * **Support Investment Decisions**: Help car dealers and investors make informed decisions based on potential returns across different car market segments.
3. * **Simplify the Buying and Selling Process**: A reliable and accessible price estimation service reduces uncertainty, enhancing the transaction experience for both buyers and sellers.

## Business Impact

The service has a significant business impact:

- * **Streamlined Decision-Making**: Accurate price estimates offer dealers a foundation for market-reflective pricing, accelerating the sales process.
- * **Enhanced Market Positioning**: A user-friendly and accurate price analysis builds customer trust and gives dealers a competitive edge.

## Comparison with Existing Solutions

Existing solutions often require extensive registration or provide only general estimates. This service delivers detailed price estimates based on specific car attributes without complex requirements, making it more accessible.

## Manual Method for Price Estimation:
Without machine learning, price estimation would require manually comparing similar cars and historical sales prices, which is time-consuming and less precise than an automated ML model.

## Business Metrics for Performance Measurement

To ensure the system meets business objectives, we track these metrics:

1. * **Price Estimate Accuracy**: Estimates should be within ±10% of actual sales prices in at least 80% of cases.
2. * **User Satisfaction**: The service should receive an average rating of at least 4 out of 5 stars.
3. * **Completed User Sessions**: At least 70% of users should complete the price estimation process.
4. * **System Availability**: The service should maintain an uptime of at least 99%.

## Machine Learning and Software Metrics

The following technical metrics measure system performance:

1. * **Accuracy**: Proportion of price estimates within ±10% of actual values, measuring model precision.
2. * **Mean Squared Error (MSE)**: Average of squared differences between predicted and actual car prices.
3. * **Latency**: Response time from user input to returned price estimate.
4. * **Throughput**: Number of price estimates handled per minute.

## Stakeholders

Key stakeholders include:

1. * **Customers**: Private individuals and professionals seeking reliable used car price estimates.
2. * **Car Dealers**: Professional users who need a tool for more accurate price assessments.

## Resources

- * **Personnel**:
  Developers for data collection and model development.
- * **Computational Resources**:
  Workstations for development and cloud resources for model execution if needed.
- * **Data Resources**:
  Historical car prices and market trends.

## Data

This project uses training and test datasets with the necessary variables for accurate price estimation. To ensure consistency and data quality, cross-validation is applied, and no sensitive personal information is collected.

## Data Preprocessing:

- * **Missing Data Handling**: Filled missing values in `fuel_type`, `accident`, and `clean_title` with inferred or default values. For numeric columns such as `horsepower`, `displacement`, and `cylinders`, iterative and simple imputers were used.
- * **Feature Engineering**: Extracted `horsepower`, `displacement`, `engine_type`, and `cylinders` from the `engine`column and encoded categorical variables, such as `brand` and `fuel_type`, using Label Encoding.
- * **Scaling**: Standardized continuous variables for improved model performance.

## Modeling

## Exploratory Data Analysis:
A correlation heatmap and price distribution plots were used to identify feature relationships. High correlations between `model_year`, `mileage`, and variables like `accident` and `price` indicated logical connections between vehicle age, mileage, and price.

## Model Selection and Optimization:
Random Forest Regressor was chosen as the primary model due to its performance with structured data. After tuning with `RandomizedSearchCV`, the optimized parameters achieved an RMSE of approximately 66,035.47 on the test set.

## Label Encoding:
We used Label Encoding rather than One-Hot Encoding to avoid the high dimensionality that One-Hot Encoding introduces. Random Forest models are effective with label-encoded data, which simplified the data representation without compromising performance.

## Deployment

The model will be deployed via Gradio, providing a user-friendly interface for car price estimates. The goal is to maintain a response time under five seconds.

Data is not stored, simplifying privacy and security compliance.

## Summary

This car price estimation project utilized Random Forest Regressor as the primary model, achieving an RMSE of 66,035.47 after optimization. Label Encoding was preferred over One-Hot Encoding to maintain a compact dataset, reducing dimensionality without compromising accuracy. The service offers accessible, reliable price estimates for private users and dealers, addressing key business goals such as optimized pricing and streamlined decision-making. The final deployment through Gradio ensures a user-friendly experience with minimal latency.

## References

Data was sourced from GitHub, and tools like Google Colab, Kaggle, and Scikit-learn supported data handling and model development. The Random Forest model selection was based on course curriculum and academic resources. Privacy regulations, including GDPR, are observed since no personal data is collected.