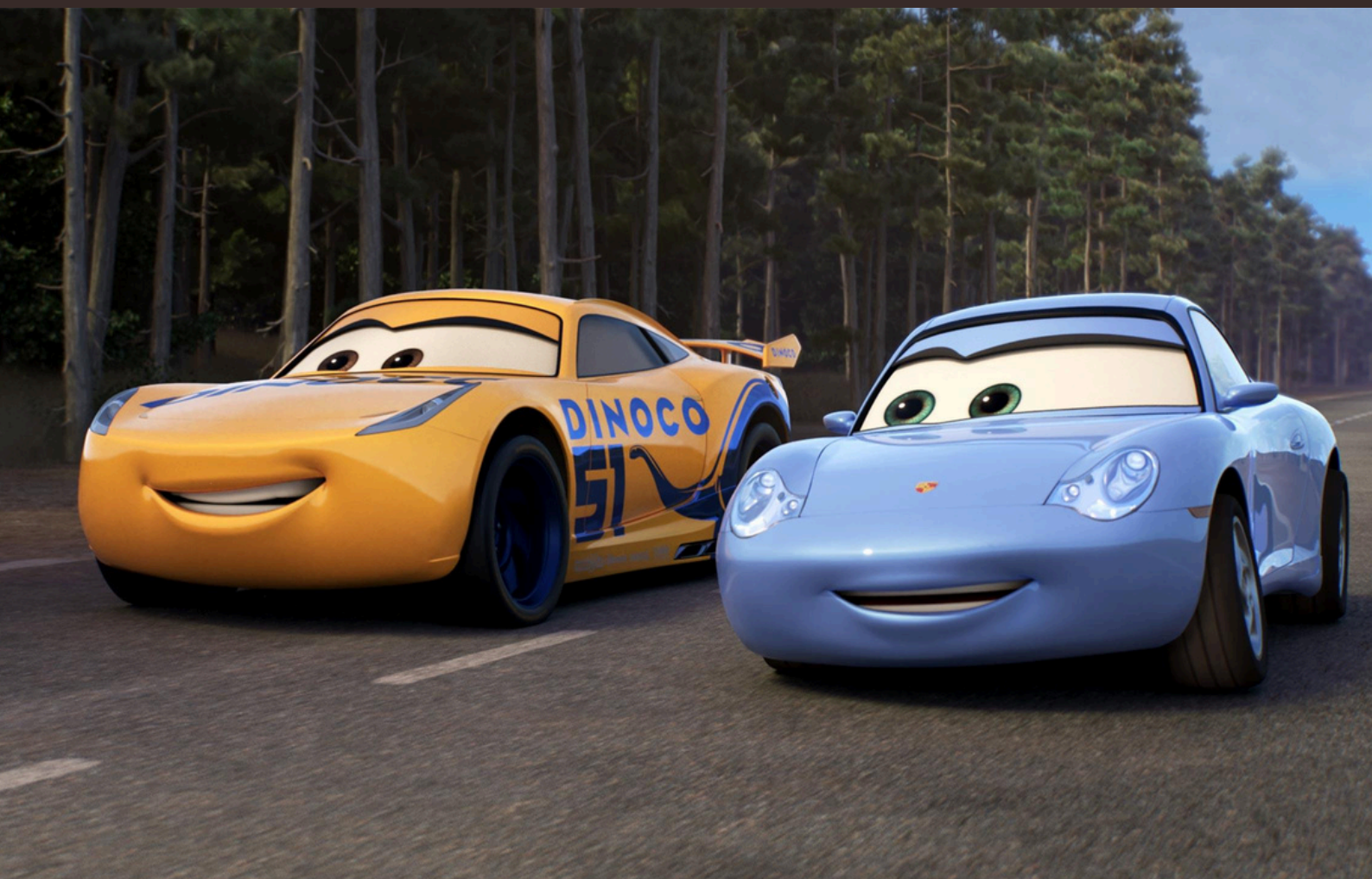
September, 2025

Cars 4 You: Expediting Car Evaluations with ML

Group Project
Machine Learning 2024/2025
Due: December 22nd (17:59)



01

I. Introduction

Cars 4 You is an online car resale company that sells cars from multiple different brands. Their business model involves an online platform in which users who which to sell their cars to provide different sets of details about the car and sending them to their chain of mechanics to get the car be evaluated before purchasing it to, later, resell the car on a profit. Using this system, the managers were able to gather an extensive list of happy customers. However, the company's growth has also led to increasing waiting lists for car inspection, which is driving potential costumers to their competitors.
To address this, the company has reached out to you. Their main goal is to expedite the evaluation process by creating a predictive model capable of evaluating the price of a car based on the user's input without needing the car to be taken to a mechanic.

II. Project Goals

The goal of your project is three-fold:

1. Regression Benchmarking: Develop a regression model that accurately predicts car prices (price). To do that, Cars 4 You has provided subset of an older version of their car database from 2020 to create a proof-of-concept. You will need to develop a consistent model assessment strategy that allows you to create and compare different candidate models to identify the most generalizable one.
2. Model Optimization: During your selection of best (or set of best) model(s), you are encouraged to explore ways to improve their performance (e.g. hyper-parameter tuning or pre-processing/feature selection adjustments). Compare the optimized model with your previous models and discuss your findings.
3. Additional Insights: This project segment is open-ended, meaning you can explore as many ideas as you desire (as long as you make them explicit and understandable). Here are some possible suggestions:
   a. Analyze and discuss the importance of the features for the different values of the target variable and how they contribute towards the prediction.
   b. Ablation Study: Measure the contribution of each element of the pipeline.
   c. Create an analytics interface that returns a prediction when new input data is provided.
   d. Test whether the best performance is achieved using a general model (trained using data from all brands/models, etc...) or using brand, model, fuel type, etc...-specific models.
   e. Determine whether training a Deep Learning network from scratch is more effective than fine-tuning a pre-trained model.

a) explainable model

b, d)improvement of scaler, imputation method, variable selection ect

e)tensorflow

02 III. Dataset

You have access to two different datasets:

In the training set, you will find the cars that were present in the Cars 4 You database in 2020. You will have features and a specific ground truth (price) associated with each assembled claim. Use the training data and the features available at the moment the user fills the form, you need to build and validate your machine-learning models.

In the test dataset, you can still access the same descriptive attributes from an independent set of cars. You must use the models trained with the training set to predict Price for all observations in the this dataset. You will not have access to the prices of the cars on this set, but you will be able to know how well your model performs on this data through a Kaggle competition.
The available data contains the following attributes:

Attribute	Description
carID	An attribute that contains an identifier for each car.
Brand	The car's main brand (e.g. Ford, Toyota)
model	The car model
year	The year of Registration of the Car
mileage	The total reported distance travelled by the car (in miles)
tax	The amount of road tax (in £) that, in 2020, was applicable to the car in question.
fuelType	Type of Fuel used by the car (Diesel, Petrol, Hybrid, Electric)
mpg	Average Miles per Gallon
engineSize	Size of Engine in liters (Cubic Decimeters)
paintQuality%	The mechanic's assessment of the cars' overall paint quality and hull integrity (filled by the mechanic during evaluation).
previousOwners	Number of previous registered owners of the vehicle.
hasDamage	Boolean marker filled by the seller at the time of registration stating whether the car is damaged or not.
price	The car's price when purchased by Cars 4 You (in £).

03

IV. Outline

Your project deliverables (notebook or zip of notebooks) should respect the following outline:

Group Member Contribution

What part(s) of the work were done by each member and an estimated % contribution of each member towards the final work.

Abstract

A small summary of your work (200 to 300 words). The abstract should give an overview of your work: What is the context? What are your goals? What did you do? What were your main results, and what conclusions did you draw from them?

I. Identifying Business Needs

- Overview and main goals of the project
- Description of the overall process and identification of model assessment approach adopted in the work (CV, LOO, Holdout, etc...)

II. Data Exploration and Preprocessing

- Description of data received -> key insights
- Steps taken to clean and prepare the data based on exploration

III. Regression Benchmarking

- Explanation of model assessment strategy and metrics used
- Feature Selection Strategy and results
- Optimization efforts: presentation, results and discussion
- Comparison of performance between candidate models

IV. Open-Ended Section

- Objectives for the Section and description of the actions taken
- Results and discussion of main findings → key takeaways

Note: This section expects that the objectives set go beyond what would reasonably be considered as adding or removing techniques to your pipeline. (e.g., using a feature selection technique not covered in class on your regular pipeline is not sufficient, but explicitly comparing different feature sets would be).

V. Deployment

The final section of your work should implement the pipeline to generate reliable predictions for new data. The output should be the .csv file that you consider the solution you selected on Kaggle as your best.

04

V. Deliverables

Upon the project's deadline, you will be required to submit:

- A clean Jupyter notebook (or a zipfile) featuring all the code you used throughout the project to:
  a. Decide on your final solution
  b. Obtain your final results (code that helped you make decisions, but does not directly contribute to reaching the goal, should be included, but commented).

VI. Evaluation

Your work will be evaluated according to the following criteria:

Criteria	Percentage (%)	Maximum Grade (out of 20)
Project Structure and Notebook(s) Quality	20	4
Data Exploration & Initial Preprocessing	20	4
Regression Benchmarking and Optimization	35	7
Open-Ended Section	20	4
Deployment	5	1
Extra Point: Have Project Be Publicly Available on GitHub	-	1

05

Your grade will reflect our assessment of the quality of your work in terms of quality of writing, clarity, conciseness, correctness and efficiency. Please find below more details about what is taken into account for each topic:

- Project Structure and Notebook(s) Quality (4v): Your notebook(s) should be readable and understandable to someone reading and looking at them for the first time. Every cell you present should have a clear purpose. This section also encompasses the overall quality of your introduction and conclusions for each cell.
- Data Exploration & Initial Preprocessing (4v): Describe the data and extract meaningful insights that you consider helpful. Avoid adding visualizations and elements that do not address the problem at hand. Additionally, it should clearly explain the steps and rationale behind the data cleaning process to make the data usable by your predictive models.
- Regression Benchmarking and Optimization (7v): Describe your strategy for model assessment. This section is separated into different components:
  - Kaggle Performance: 1v
  - Additional Preprocessing: 0.5v
  - Feature Selection Strategy and Implementation: 1.5v
  - Modelling approach - model assessment strategy (holdout, cross-validation, etc...) and algorithms (minimum of 5 covered in class) used: 2v
  - Performance assessment - rationale for choice of evaluation metric(s) and interpretation of results: 1.5v
  - Model optimization: 1v
- Open-Ended Section (4v): Describe your strategy for the additional insights objective. This section is separated into different components:
  - Formulation and Adequacy of the Objectives: 0.5v
  - Difficulty of tasks: 1v
  - Correctness/efficiency of implementation: 1v
  - Discussion of results: 1v
  - Alignment between results and communicated objectives: 0.5v
- Deployment (1v): In the final part of the project, you should be able to take your test data and ensure that it follows the steps implemented in the previous stages of the project.

06 VII. Parting Notes

1. Deliveries made after the deadline will incur a penalty of 1 point per day. Deliveries made before the deadline will receive a bonus of 0.15 points per day of delivery in advance (up to a maximum of 1 point).
2. If you use automated packages (e.g., ydata, pandas-profiling), we will only consider observations that have a visible impact on the project (i.e., a key insight that is acted upon).
3. When it comes to creating Predictive Models, you are allowed to create your own implementations or use one of the following packages: vanilla Scikit-Learn, Tensorflow, Keras, PyTorch (and skorch) or Transformers. optuna for optimisation
4. Using Lazy Predict or similar AutoML (e.g., Feature Tools, TSFresh) packages, XGBoost, CatBoost, LightGBM (or similar packages not in vanilla sk-learn) is explicitly off-limits and will result in a 1-point penalty per attempt.
5. We won't consider any steps or results that are not mentioned in any of your submitted notebooks.
6. Everything in your clean notebook must have a clear purpose. Avoid including irrelevant, unimportant, or redundant information. Additionally, please refrain from providing theoretical explanations of topics covered in class.
7. The trustworthiness of the information you provide is key. You should look to source information from peer-reviewed journals (thus, avoid citing Medium, TowardsDataScience and similar sources). Also, for the love of God, DO NOT CITE ChatGPT or any other LLM for that matter.
8. Before submitting, run your notebook from the start one last time (if you used a GridSearch, you can comment this cell, but you should run the final model with the GS parameters in a different cell).
9. If you tested more elements that should be considered, but did not feature in your clean notebook, please submit the code you used in a separate submission form created for that purpose.
10. We will run your Jupyter Notebooks. Please ensure that we can run the notebook from start to finish without interruption. Notebooks that do not fulfil this condition will be severely penalized.
11. The code will pass through a process of plagiarism and AI generation checking.
12. You must submit to the Kaggle competition to get points for that component.
13. When determining the grade for your work, there will be a comparative component between it and the work presented by your peers.

Friendly Reminders:

1. The order in which the group members present their notebooks is decided by the professors at the start of the discussion. Not attending will result in a 0.
2. If something is good enough to be mentioned, it is also good enough to know. DO NOT include techniques/algorithms/steps you cannot explain in your work: we may (and probably will) ask about them in the discussion.
3. Finished is better than perfect.