About Course

Foundations of Intelligent Systems: Algorithms

Name

About Submission

Car Prices Prediction

Brian Maina Nyawira

27th June 2025

Problem definition

The core problem is to accurately predict the **sales price of used cars based on their attributes**.

Solution: Building a **machine learning model** that can **learn complex relationships between various car characteristics** (such as make, model, year, mileage, engine type, transmission, body style, etc.) and their corresponding market prices.

Justification

- **Buyers** when purchasing a used car, consumers often lack transparent and reliable pricing information.
- **Sellers (Individual & Dealerships)**: For individuals selling their car: knowing its true market value is crucial for setting a competitive price, attracting buyers, and maximizing profit.
- **For Automotive Businesses:** Dealerships can optimize their used car inventory by predicting which cars will sell quickly at what price.
- **Risk Assessment:** Financial institutions and lenders providing car loans can use these predictions for better risk assessment, ensuring the loan amount aligns with the car's actual value.
- **Marketing:** This data can inform manufacturing strategies and marketing campaigns.

About the Dataset

Link to Dataset https://www.kaggle.com/datasets/nelgiriyewithana/australian-vehicle-prices

The Australian vehicle prices (2023) dataset has **19** columns and **16733** rows of data:

Missing Values

Brand	1	Engine	1	Location	450
Year	1	DriveType	1	CylindersinEngine	1
Model	1	FuelType	1	BodyType	282
Car/Suv	28	FuelConsumption	1	Doors	1604
Title	1	Kilometres	1	Seats	1705
UsedOrNew	1	ColourExtInt	1	Price	3
Transmission	1				

Key feature is the prices and how they are affected by Year, Kilometers, Engine-Litres, Fuel Consumption and a combination of the other features.

Numerical: `Year, Kilometres, Price. FuelConsumption`
Categorical: `Brand, Model, Car/Suv, Title, UsedOrNew, Transmission, Engine, DriveType, FuelType, ColourExtInt, Location, CylindersinEngine, BodyType, Doors, Seats, Engine_Cylinders.`

Pre-Processing

For the `prices` column I replaced POA, $, with ' ' and dropped all null values. The `year` was a float e.g 2020.0 and extracted the four digits of the year. For all missing values in `Doors, Fuel consumption, Kilometers and Seats` I replaced with the median value. In case of values **>=10** for both `doors and seats` I handled them to cater for data entry error. I dropped problematic column `CAR/SUV` and instead used `BodyType`. Reduced `Model` cardinality to top 350 models + `Other_Model`. Dropped `Title` column. I extracted `Interior_Material` and `Interior_Color` from `ColourExtInt`.

	Brand	Year	Model	Car/Suv
0	Ssangyong	2022.0	Rexton	Sutherland Isuzu Ute
1	MG	2022.0	MG3	Hatchback
2	BMW	2022.0	430I	Coupe
3	Mercedes-Benz	2011.0	E500	Coupe
4	Renault	2022.0	Arkana	SUV

	Title	UsedOrNew	Transmission
0	2022 Ssangyong Rexton Ultimate (awd)	DEMO	Automatic
1	2022 MG MG3 Auto Excite (with Navigation)	USED	Automatic
2	2022 BMW 430I M Sport	USED	Automatic
3	2011 Mercedes-Benz E500 Elegance	USED	Automatic
4	2022 Renault Arkana Intens	USED	Automatic

	Engine	DriveType	FuelType	FuelConsumption	Kilometres	ColourExtInt
0	4 cyl, 2.2 L	AWD	Diesel	8.7 L / 100 km	5595	White / Black
1	4 cyl, 1.5 L	Front	Premium	6.7 L / 100 km	16	Black / Black
2	4 cyl, 2 L	Rear	Premium	6.6 L / 100 km	8472	Grey / White
3	8 cyl, 5.5 L	Rear	Premium	11 L / 100 km	136517	White / Brown
4	4 cyl, 1.3 L	Front	Unleaded	6 L / 100 km	1035	Grey / Black

	Location	CylindersinEngine	BodyType	Doors	Seats	Price
0	Caringbah, NSW	4 cyl	SUV	4 Doors	7 Seats	51990
1	Brookvale, NSW	4 cyl	Hatchback	5 Doors	5 Seats	19990
2	Sylvania, NSW	4 cyl	Coupe	2 Doors	4 Seats	108988
3	Mount Druitt, NSW	8 cyl	Coupe	2 Doors	4 Seats	32990
4	Castle Hill, NSW	4 cyl	SUV	4 Doors	5 Seats	34990

Dataset head

Methodology

Tools and Libraries

Data manipulation: pandas, and numpy,
Data visualization: matplotlib, and seaborn,
Model selection and evaluation: sklearn.model_selection, and sklearn.metrics,
Ensemble methods: RandomForestRegressor, and xgboost
Preprocessing: StandardScaler,
Hyperparameter tuning: RandomizedSearchCV,
Feature selection SelectFromModel,
Model persistence: joblib

Scaling: use standardization scaling when we desire faster convergence [1]

[2]normalization techniques do not handle the outlier problem as effectively as standardization because standardization explicitly relies on both the mean and the standard deviation

Reducing the Curse of Dimensionality: using feature selection using embedded methods[3] rather than PCA as interpretability of features is needed. Selected Features is **76** out of over 1000 features after one-hot encoding.

[3]Filter based method learning algorithms are not used for feature selection, whereas Wrapper based method uses the learning algorithm for testing the quality of selected feature subsets. Embedded Method overcomes the computational complexity.

Methodology

XGBOOST

XGBoost: It's an ensemble method that builds trees sequentially, with each new tree correcting the errors of the previous ones. It includes built-in regularization, handles missing values, and is highly optimized.

[4]The detailed analysis of these two machine learning algorithms concludes that XGBoost has the upper hand over Random Forest in multiple dimensions.

Best Parameters for XGBoost: `'subsample': 1.0, 'reg_lambda': 0.001, 'reg_alpha': 1, 'n_estimators': 500, 'max_depth': 3, 'learning_rate': 0.2, 'gamma': 0, 'colsample_bytree': 0.6`

Performance:

```
Root Mean Squared Error (RMSE): 14029.18
R-squared (R2 Score): 0.8605
```

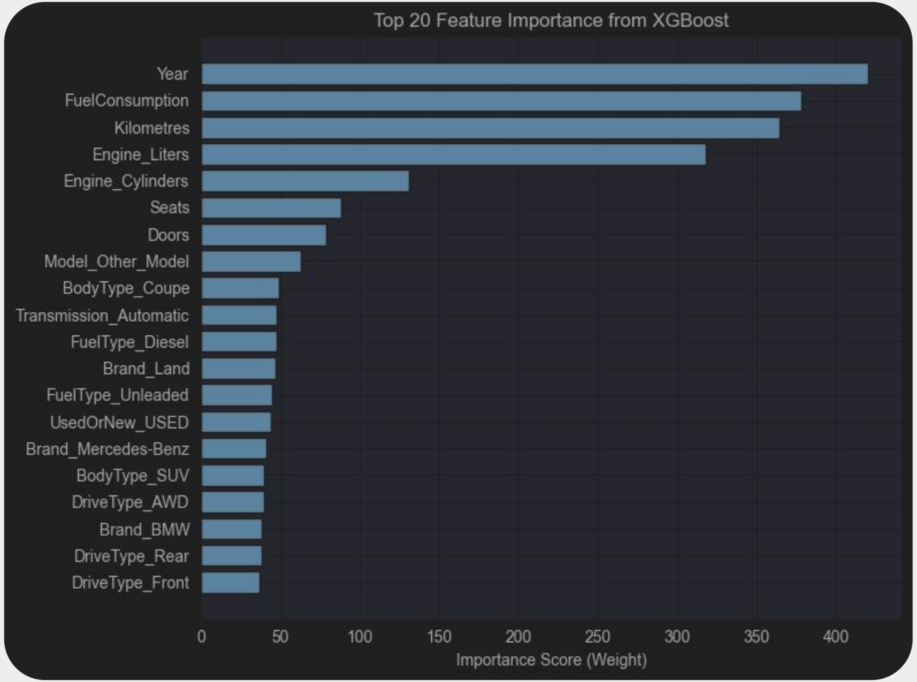
Results and Analysis

Visualization

Results and Analysis

Visualization

Results and Analysis



Top Predictors are as Expected: `Year, Kilometres, Engine_Liters, and Engine_Cylinders` are foundational, which aligns perfectly with intuition and earlier correlation analysis.

Market Nuances Captured: The importance of features like `FuelConsumption, specific Brand_ names, Model_Other_Model, Transmission_Automatic, and BodyType_SUV` demonstrates that XGBoost model captured sophisticated market preferences and segmentations.

Value of One-Hot Encoding: The high importance of the one-hot encoded features validates the extensive preprocessing efforts. These categorical distinctions are meaningful to the model.

Challenges and Solutions

The column `Car/Suv` contains a mix of car types and dealership names and I ended up dropping it

The column `Title` was more of description of the car; I had to drop it as it bore features found also in the other columns.

The columns `Interior_Material` and `Interior_Color` were not explicitly stated but part of the `ColourExtInt` column thus by extracting them from it a majority of the rows didn't have the two features.

Future Work

Try other powerful ensemble methods like LightGBM or CatBoost, which are often competitive with XGBoost and can sometimes be faster

References

1. Sharma, V. (2022). A Study on Data Scaling Methods for Machine Learning. *International Journal for Global Academic & Scientific Research*, 1(1). doi:https://doi.org/10.55938/ijgasr.v1i1.4.
2. Shaibu, S. (2024). *Normalization vs. Standardization: How to Know the Difference*. [online] Datacamp.com. Available at: https://www.datacamp.com/tutorial/normalization-vs-standardization.
3. Venkatesh, B. and Anuradha, J. (2019). A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies*, [online] 19(1), pp.3–26. doi:https://doi.org/10.2478/cait-2019-0001.
4. Fatima, S., Hussain, A., Sohaib Bin Amir and Syed Haseeb Ahmed (2023). XGBoost and Random Forest Algorithms: An in Depth Analysis. *Pakistan journal of scientific research*, 3(1), pp.26–31. doi:https://doi.org/10.57041/pjosr.v3i1.946.

Q&A

Questions and Answers

Conclusion

Thank you