School of Computing
FACULTY OF ENGINEERING AND
PHYSICAL SCIENCES

UNIVERSITY OF LEEDS

Final Report

Machine Learning Methods to Estimating Used Cars’ Prices

Gengchen Han

**Submitted in accordance with the requirements for the degree of
BSc Computer Science**

2023/24

COMP393 Individual Project

The candidate confirms that the following have been submitted*:*

Items	**Format**	**Recipient(s) and Date**
Final Report	*PDF file*	*Uploaded to Minerva (01/05/24)*
Link to online code repository	*URL*	*Sent to supervisor and assessor (01/05/24)*

The candidate confirms that the work submitted is their own and the appropriate credit has been given where reference has been made to the work of others.

I understand that failure to attribute material which is obtained from another source may be considered as plagiarism.

(Signature of student) Gengchen Han

Summary

In the current used car market, accurately predicting the value of a vehicle is of great significance to buyers, sellers, and market analysts. With the development of big data and machine learning technology, it is possible to use various characteristics of a car to predict its second-hand price. The purpose of this study is to develop a predictive model to estimate the market value of used cars by analyzing a variety of vehicle characteristics such as make, model, fuel type, body style, drive wheel type, engine size, horsepower and other factors.

By collecting and analyzing a dataset containing a variety of vehicle characteristics, the study uses a variety of data pre-processing techniques to prepare the data, and then applies a variety of machine learning algorithms, such as linear regression, decision trees, random forests, and gradient elevators, to build and train the predictive model. By comparing the performance of these models on the test data, a model with the best performance is finally selected to forecast the used car price.

In short, this research has successfully developed a used-car price prediction model based on vehicle characteristics through the application of machine learning technology, and hopes that this research can help participants in the used-car market - both sellers and buyers - to make more intelligent decisions, and at the same time contribute to the transparency and standardization of the used-car market.

Acknowledgements

Firstly, I would like to thank my supervisor, Natasha. She has been a great help to me in terms of her expertise and skills. Her expertise, meticulous guidance and spirit of encouragement have greatly contributed to my academic growth and the success of my project.

Secondly, I would like to thank my personal tutor, Mr. Martin, who gave me great support and guidance when I encountered difficulties in my life.

Finally, I must thank my parents. Their support and love for me in all aspects is a strong backing for me to successfully complete my study and graduation project.

Table of Contents

Chapter 1 Background Research

1.1 Used car situation

In today's society, the car is no longer a luxury, but has become an important tool for many People's Daily commute and life. As new car prices continue to rise, more and more consumers are turning to the used car market for more cost-effective options. The booming of the used car market not only provides consumers with more choices, but also promotes the recycling of resources, which has important economic and environmental significance.

The global used car market has experienced significant growth in recent years and is expected to continue to expand. The global used car market was valued at $1.66 trillion in 2022 and is expected to continue growing at a CAGR of about 6.1% from 2023 to 2030. (as shown below)[1]

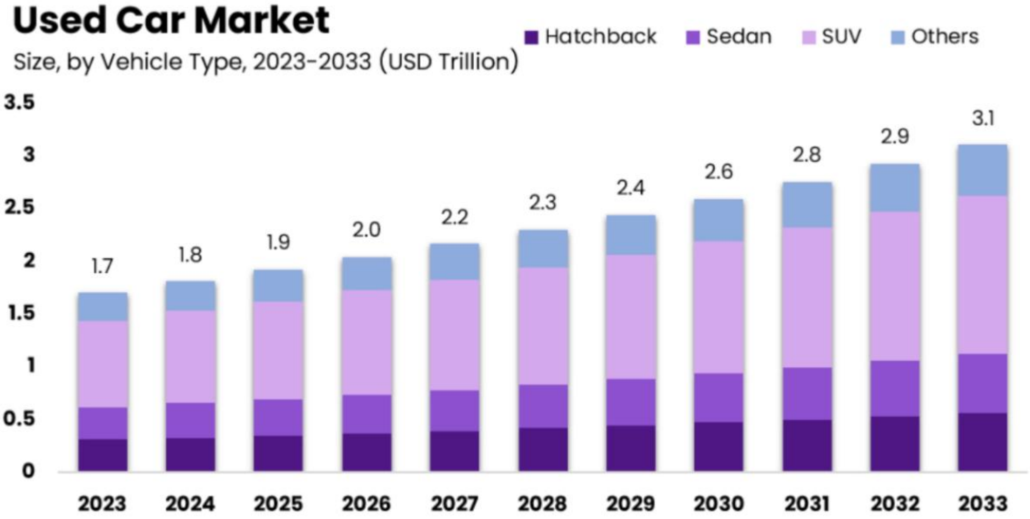


Figure 1: Forecast of used car sales from 2023 to 2030

As technology advances, data analytics and machine learning provide new tools to solve this conundrum. We can train computer models to predict used car values by collecting data on historically traded used cars and analyzing their characteristics and transaction prices. In this way, we expect to provide a faster and more objective method of valuing used cars.

1.2 Critical analysis of existing solutions and techniques

Most consumers do not know much about the price of used cars, so they will listen to the opinions of second-hand car dealers to a large extent when buying, which may lead to consumers spending much higher than the normal price for second-hand cars.

There are large websites such as Carfax that provide a detailed overview of used car history, including accident records, repair history and ownership changes. These reports are a valuable source of information for buyers, but they do not provide price forecasts and, with incomplete data records, may not fully reflect the true condition of the vehicle.

Online platforms such as AutoTrader[2] and Cars.com[3] help users understand vehicle values by bringing together vast amounts of real-time market data. These platforms match user input vehicle information with similar vehicles for sale to provide a price reference. Still, these prices are artificially equalized, and there are bound to be many sales that are above or below the value of the used car that should really be sold.

Recently, machine learning technology has been applied to the prediction of used car prices. These techniques can analyze a large number of data points and learn how different characteristics, such as year and mileage, affect prices. However, the accuracy of AI models largely depends on the quality of the training data. If the input data is biased or incomplete, the forecast results may also be inaccurate.

1.3 Description of software prototypes

Data Preprocessing: The software will include a data preprocessing module that can clean and format input data, identify and process missing or outliers, and ensure that subsequent models receive high-quality data.

Analysis model: The prototype will integrate multiple machine learning algorithms such as random forest, gradient elevator, neural network, etc., so that users can choose the most suitable algorithm according to their needs. Each algorithm will be trained on historical transaction data, learning the relationship between different features and prices.

1.4 Requirements and risk analysis

Demand analysis:

Accuracy: All stakeholders need models that can provide accurate price predictions so that both buyers and sellers can get a fair deal.

Transparency: For buyers in particular, the basis of the forecast must be transparent and they need to understand how the price forecast is arrived at.

Real-time updates: Market prices fluctuate frequently, and models must be able to reflect real-time market conditions.

Reliability: As the market expands, the model should be able to easily scale to accommodate more data and complex analysis.

Risk analysis:

Data quality: Model accuracy is highly dependent on the quality of the training data. Inaccurate or incomplete data may lead to misleading price forecasts.

Technological adaptability: With the advent of new technologies, existing predictive models can quickly become obsolete.

Legal risks: Models need to comply with various legal and industry standards, and non-compliance can pose legal risks.

1.5 existing theoretical techniques

Linear Regression[4]:

As a classical forecasting model, the core idea of linear regression is to assume that there is a linear dependence between the target variable (dependent variable) and a series of independent variables. In the context of used car price forecasting, linear regression models are particularly useful when the relationship between the price of a used car and its characteristics (e.g., year of manufacture, miles driven, vehicle condition, etc.) can be roughly described as linear. By fitting data points to a straight line (or hyperplane in multiple linear regression), the linear regression model quantifies the impact of features on price and makes predictions about used cars at unknown prices. Linear regression model is not only easy to explain and understand, but also can provide stable and reliable prediction results under the condition that the data meet the linear hypothesis.

Decision Tree Regressor[5]:

Decision tree regression is a predictive model based on tree structure that learns decision rules in the data by recursively partitioning the data set into subsets. In the decision tree, each "node" represents the decision point of a feature, divides the data into different subsets based on different values of the feature, and finally gives the prediction result through the leaf node. For the used car price prediction problem, decision tree regression performs well when dealing with data containing classified features (such as car brand, model, etc.). Because the structure of decision tree is intuitive and easy to understand, it can clearly reveal the hierarchical relationship between features and prices, and provide an intuitive basis for decision making. In addition, decision tree regression has strong robustness to outliers and missing values, and can maintain good prediction performance in complex data environments. However, it is worth noting that decision tree regression may be at risk of overfitting, so overfitting needs to be avoided through techniques such as appropriate pruning when applied.

Random Forest Regressor[6]:

Random forest regression is a powerful prediction algorithm based on ensemble learning, which builds and integrates the results of multiple decision trees to improve the accuracy and stability of predictions. The core idea is to use bootstrap method to randomly select multiple subsets from the original data set and build decision trees on each subset. During the construction of these decision trees, feature subsets are randomly selected for splitting, thus introducing some randomness. By averaging the prediction results of all decision trees (for regression problems), random forest regression can effectively handle complex relationships between features, including nonlinear relationships, and show good robustness in the face of outliers and noisy data. In the field of used car price prediction, random forest regression can comprehensively consider the influence of multiple features on price, and give accurate and reliable prediction results.

Gradient Boosting Regressor[7]:

Stepwise regression is a powerful technique for building predictive models iteratively by combining multiple weak predictive models (such as simple decision trees) into a powerful integrated model. In each iteration, gradient-lifting regression trains a new weak prediction model based on the residual (prediction error) of the previous iteration, and the predictions from that model are added to the total prediction. In this way, gradient lifting regression can gradually improve the prediction results, reduce the prediction errors, and finally obtain a high-precision prediction model. Because of its ability to deal with nonlinear relationships, automatically select features and adapt to complex data distribution, gradient lifting regression shows excellent performance in practical problems such as used car price prediction.

Support Vector Regression (SVR)[8] :

Support vector regression (SVR) is the application of support vector machine (SVM) to regression problems, and it is also a powerful prediction algorithm. The goal of SVR is to find an optimal decision boundary (hyperplane) that maximizes the distance from the data point to the hyperplane while minimizing the prediction error. Similar to SVMS in classification problems, SVR can also handle nonlinear problems by using different kernel functions such as linear kernel, polynomial kernel, radial basis function kernel, etc. In the used car price prediction, SVR can capture the complex relationship between features and give accurate prediction results. In addition, SVR is insensitive to the size and dimension of the data, is suitable for data sets of all sizes, and can still maintain good performance when processing high-dimensional data.

Chapter 2 Methods

2.1 Scheme design and software architecture

2.1.1 Problem Definition

In today's dynamic and competitive automotive market, accurately predicting used car prices is critical for buyers, sellers, and market analysts alike. Price forecasting not only helps consumers make informed purchasing decisions, but also allows sellers to determine reasonable selling prices, thereby increasing sales opportunities and profitability.

Although a variety of valuation tools and services exist in the market, they are often limited by fixed valuation models and historical data, which are difficult to adapt to the rapid changes in the market and the diversity of vehicle characteristics. In addition, these tools often lack transparency, making it difficult for users to understand and trust the predicted results. The model accuracy of the research on this problem on the Internet is about 78%, and I decided to improve the accuracy of this model through the model I built.

Therefore, the goal of this project is to develop a machine learning-based used car price prediction model that takes into account various vehicle characteristics such as make, model, vehicle age, mileage, vehicle condition, etc., and updates the data in real time to provide more accurate and reliable price predictions. We expect this model to bring significant economic benefits and decision support to all participants in the used car market.

2.1.2 Data overview

The dataset used in this study is derived from the Kaggle platform, which aims to predict the market price of used cars through a series of detailed car attributes. The data set contains 19,237 used car transaction records, each consisting of 18 characteristics, covering the vehicle's make, model, year of production, age, mileage, engine size, number of doors, number of airbags and whether it has a turbo. These characteristics reflect the overall condition and performance of the vehicle and are critical to predicting its market value.

2.1.3 Feature model selection

First, data quality and integrity is one of the major challenges affecting model accuracy. First, I used EDA technology to visualize and analyze the data, observe the distribution, trend and relationship pattern of the data, and understand the structure and dynamics of the data. Observe whether there is correlation between variables, whether the distribution is normal, etc. Identify which features are likely to be useful for predictive models and which may be redundant or irrelevant.

Secondly, the complexity of feature selection is also a difficult point that cannot be ignored. The price of a used car is affected by a variety of factors, including the make, model, age, mileage, and condition of the vehicle. Choosing which features to enter the model and how to deal with these features will directly affect the predictive performance of the model. To do this, I employ machine learning algorithms such as linear regression, random forest, and gradient lifting, which automatically identify important features and accurately predict models.

2.1.4 Tuning hyperparameters

After the model was first predicted, I chose a grid search technique to optimize several well-performing models, which allows us to automatically test multiple possible parameter combinations and find the best one (similar to arranging all possible parameter combinations in a grid, trying one by one, until we find the best one). This process needs to be done automatically by the computer and output the optimal combination of parameters for us to use. In this way, we can save a lot of time and effort to manually adjust the parameters, and it can be easier to find the best combination of parameters, so that our machine learning models become more accurate and reliable.

2.1.5 Model evaluation

Two statistical indicators, R2 score and RMSE (root mean square error), were used in this project.

The R2 score[9], also known as the coefficient of determination, is a measure of how well the model fits the data. Its value is usually between 0 and 1, with 1 indicating that the model fits the data perfectly. If the R2 score is negative, it means that the model is performing worse than the average model. The R2 score can be interpreted as the proportion of data variability that the model can account for.

The RMSE[10] is the square root of the mean of the squared values of the difference between the observed values and the predicted values of the model. It measures the standard deviation of the forecast error. The lower the RMSE value, the smaller the prediction error and the better the prediction performance of the model.

2.2 Version Control

This project used anaconda Jupyter Notebook for code writing, and I uploaded all the resources related to the project to github. I upload my progress to git after completing a small portion of each task. To achieve version control.

2.3 Schematic diagrams of overall software architecture.

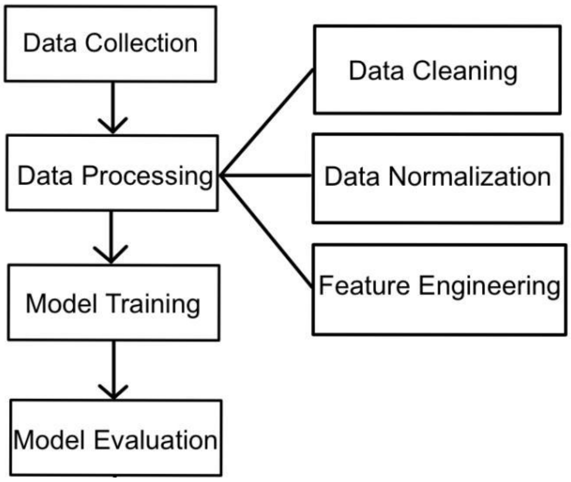


Figure 2: Schematic diagrams of overall software architecture

Data Collection:

In a used car price forecasting project, data collection is a crucial first step. In order to obtain comprehensive and accurate data, this study used the online service kaggle.com to collect multi-source data on used cars. This data may include the vehicle's make, model, year, mileage, and so on. The collected data will serve as the basis for subsequent analysis, providing a wealth of information for this study to build predictive models.

EDA (Exploratory Data Analysis) :

After data collection, exploratory data analysis (EDA) was performed to gain insight into the data. By applying visualization and basic statistical techniques, this study aims to discover key features, trends, and patterns in the data. EDA can not only help us to understand the distribution and relationship of data, but also reveal potential data anomalies and missing values, and provide guidance for subsequent data preprocessing and feature engineering.

Data Preprocessing:

This study first cleaned the dataset, removed duplicate records, dealt with missing and outliers, and ensured the integrity and accuracy of the data. Subsequently, this study standardized the data and adjusted all numerical features to the same scale to eliminate dimensional effects between features, thus improving the efficiency of model training and the reliability of prediction results. In addition, this study conduct feature engineering, by mining and creating new features, such as feature combinations derived from existing data, these efforts are aimed at enriching model inputs and improving their accuracy for price predictions.

Model Training:

This study uses a variety of algorithms to construct prediction models, including linear regression, decision tree regression, and random forest regression. These algorithms were chosen for their wide application and proven effectiveness in predicting used car prices. In order to optimize the model parameters and enhance its generalization ability in various data environments, this study adopt a strategy of cross-validation. In addition, in order to prevent the model from overfitting, this study adjusted the complexity of the model and introduced regularization parameters, thus improving the stability and robustness of the model.

Model Evaluation:

This study uses two key indicators, R^2 score (coefficient of determination) and RMSE (root mean square error), to evaluate the performance of the model. The R^2 score reflects the correlation between the predicted value of the model and the actual value, and the closer the value is to 1, the higher the prediction accuracy of the model. The RMSE index represents the average error degree between the predicted value of the model and the actual value, and the smaller the value, the higher the prediction accuracy of the model. By calculating these metrics and comparing their performance with other models, this study are able to assess the effectiveness of the selected model in predicting used car prices and select the best-performing model for subsequent application.

2.4 Management methodology

Agile development in used car price prediction project[11]

During the implementation of the used car price prediction project, the methodology of agile development was adopted to ensure the rapid iteration and continuous improvement of the project. The following are specific applications and descriptions of agile development at various stages of a project:

1. Project planning and framework determination

At the beginning of the project, the study clearly defined the objectives, scope and expected results of the project. Through the rapid iteration and continuous improvement approach of agile development, this study sets the initial direction and framework for the project. This study emphasizes on building the skeleton of the project at the initial stage to lay a solid foundation for the detailed implementation and expansion of the follow-up.

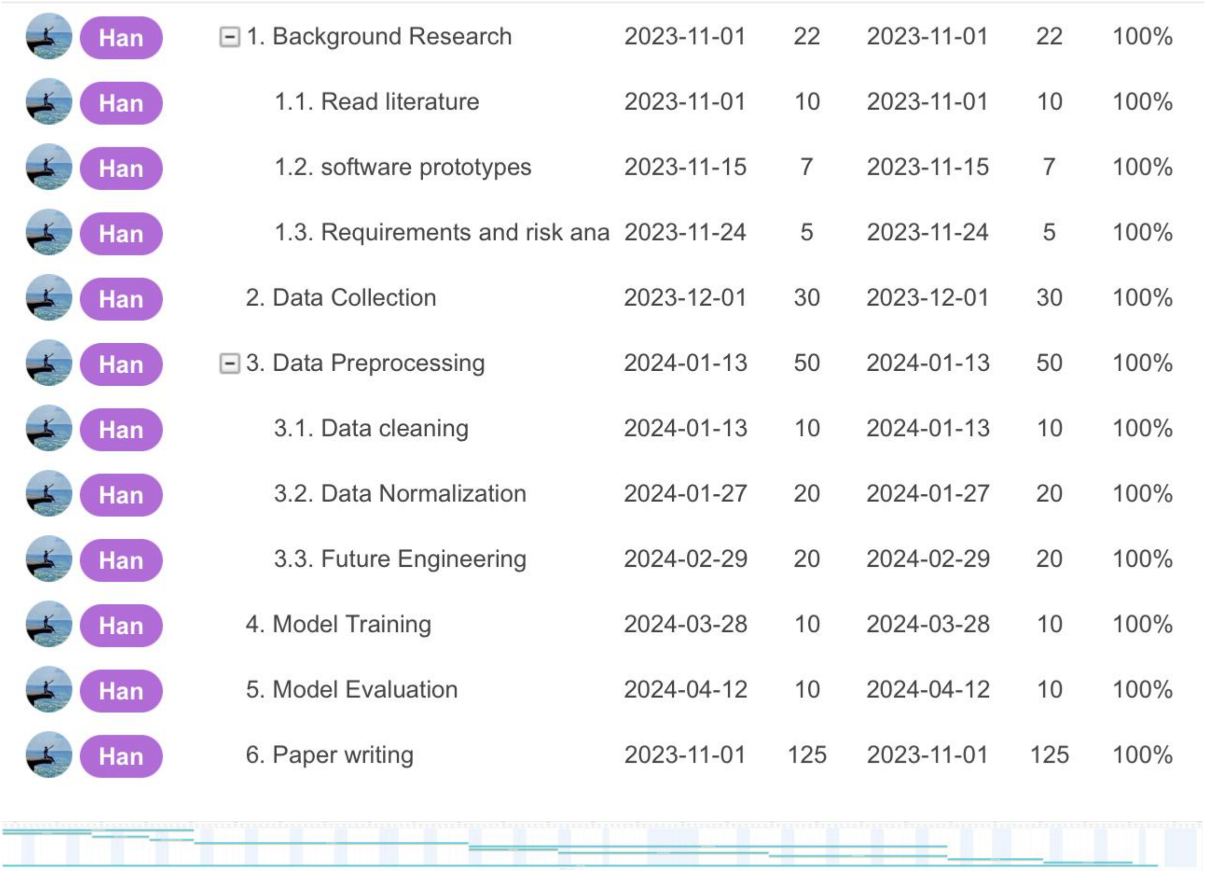
2. Demand analysis

Based on a deep understanding of the business requirements, this study conduct detailed data and feature requirements analysis. this study recognize that there are many factors that affect used car prices, including but not limited to vehicle make, model, year, mileage, vehicle condition, etc. Therefore, this study collected and analyzed a large amount of relevant data to ensure that the predictive model can fully account for these influencing factors.

3. Time management

In agile development, time management is critical. This study uses an iterative workflow that divides the project into several small, executable iterative cycles. Within each iteration cycle, the study set clear milestones and delivery points, and used tools such as Gantt charts[12] to plan and track progress. This approach gives us the flexibility to respond to changes and uncertainties in the project, ensuring that the project moves forward on time. (The Gantt chart is as follows)

The names of each column are: Person in charge, Quest, Planned Start Day, Planned day, Actual start day, and Actual day progress



Han	⊟ 1. Background Research	2023-11-01	22	2023-11-01	22	100%
Han	1.1. Read literature	2023-11-01	10	2023-11-01	10	100%
Han	1.2. software prototypes	2023-11-15	7	2023-11-15	7	100%
Han	1.3. Requirements and risk ana	2023-11-24	5	2023-11-24	5	100%
Han	2. Data Collection	2023-12-01	30	2023-12-01	30	100%
Han	⊟ 3. Data Preprocessing	2024-01-13	50	2024-01-13	50	100%
Han	3.1. Data cleaning	2024-01-13	10	2024-01-13	10	100%
Han	3.2. Data Normalization	2024-01-27	20	2024-01-27	20	100%
Han	3.3. Future Engineering	2024-02-29	20	2024-02-29	20	100%
Han	4. Model Training	2024-03-28	10	2024-03-28	10	100%
Han	5. Model Evaluation	2024-04-12	10	2024-04-12	10	100%
Han	6. Paper writing	2023-11-01	125	2023-11-01	125	100%

Figure 3:Gantt Chart

4. Quality assurance

In agile development, this study always puts quality first. In order to ensure the quality of predictive models and software, a series of quality control measures were developed in this study. This includes code reviews, model validation, and testing. This study emphasizes continuous quality checks during the development process to ensure that the results of this study meet the expected standards and requirements.[13]

5. Communication and coordination

Agile development emphasizes close communication and collaboration between teams. To ensure transparency and timely sharing of information, the study set a regular meeting schedule, including face-to-face meetings with project supervisors.

Every two weeks, I have a meeting with the supervisor to report on the progress of the project, discuss problems encountered, and seek better advice. This regular communication mechanism helps us to identify and solve problems in a timely manner and ensure the smooth progress of the project.[14]

6. Project delivery and evaluation

At the end of each iteration cycle, the project will be evaluated and summarized. This study conducts acceptance tests on the completed results to ensure that they meet the needs and achieve the desired results. At the same time, this study will also reflect and summarize the entire iteration process to find areas that can be improved and make a more optimized plan for the next iteration cycle. This continuous evaluation and improvement mechanism helps the research to continuously improve the quality and efficiency of the project.[15]

Chapter 3 Process and Result

3.1 Process

3.1.1 ETL (Extract, Transform, Load)

For the original data set, this project adopts the ETL (Extract, Transform, Load)[13] method for initial processing. ETL is a term commonly used in data warehouse technology, which is used to describe the extraction of data from multiple sources and the appropriate cleaning and transformation of data. The process of then loading into another database or data warehouse.

3.1.1.1 Extract

At this stage, the raw data is obtained from the kaggle website as a csv file under license. Then read it into the Jupyter Notebook.

3.1.1.2 Transform

The process of cleaning and integrating extracted data in order to load it into the target system.

Missing value handling: Convert all '-' values in the levy column to '0', and then convert them to floating-point type for subsequent computation.

Modify variable names: For example, the name of one column is Prod. year, which means the production date of the car. However, in order to facilitate calculation and make the data clearer, this project uses the year of this year minus the production years as the service years of the car, and changes the column name to Age.

Modify the variable type: the data in the Leather interior column was originally a string type, which would cause difficulties in later operations, so I converted the data type of the Leather interior column to bool.

Delete duplicate data: After statistics, 313 duplicate data were found, which would affect the experiment results, so they were deleted.

Comprehensive treatment:

The combination of units is added to the data in the Mileage column. In order to facilitate subsequent operations, I remove the units in this column and change the type of the data to float. It is found that more than 1000000 Mileage points may be outliers, so the whole line of data with outliers is deleted.

The data in the Engine volume column is the displacement of the engine. Some cars have turbo, so some data will be displayed as: '2.4Turbo'. This type of data will affect subsequent calculations. Therefore, this experiment deleted Turbo characters in all data with Turbo, leaving only numerical values, and then changed the data type of column names to float. Then add a new column, the effect is whether to turbo. The column name is Turbo. The type is set to bool. 1 indicates that Turbo exists, 0 indicates that turbo does not exist.

The data in the column of Doors means the number of doors owned by the car. In the original data, the number of doors was counted with strings. In this experiment, adjustments were made, by replacing 04-May with 4 of int type and 02-Mar with 2 of int type, &gt; 5 is replaced with 6 of type int to make subsequent operations easier and clearer.

Then the experiment carried out large-scale outlier processing on the data, traversed the whole data set, and made the boxplot of each variable. This makes it clear which column contains outliers (see figure below), so outliers can be removed.

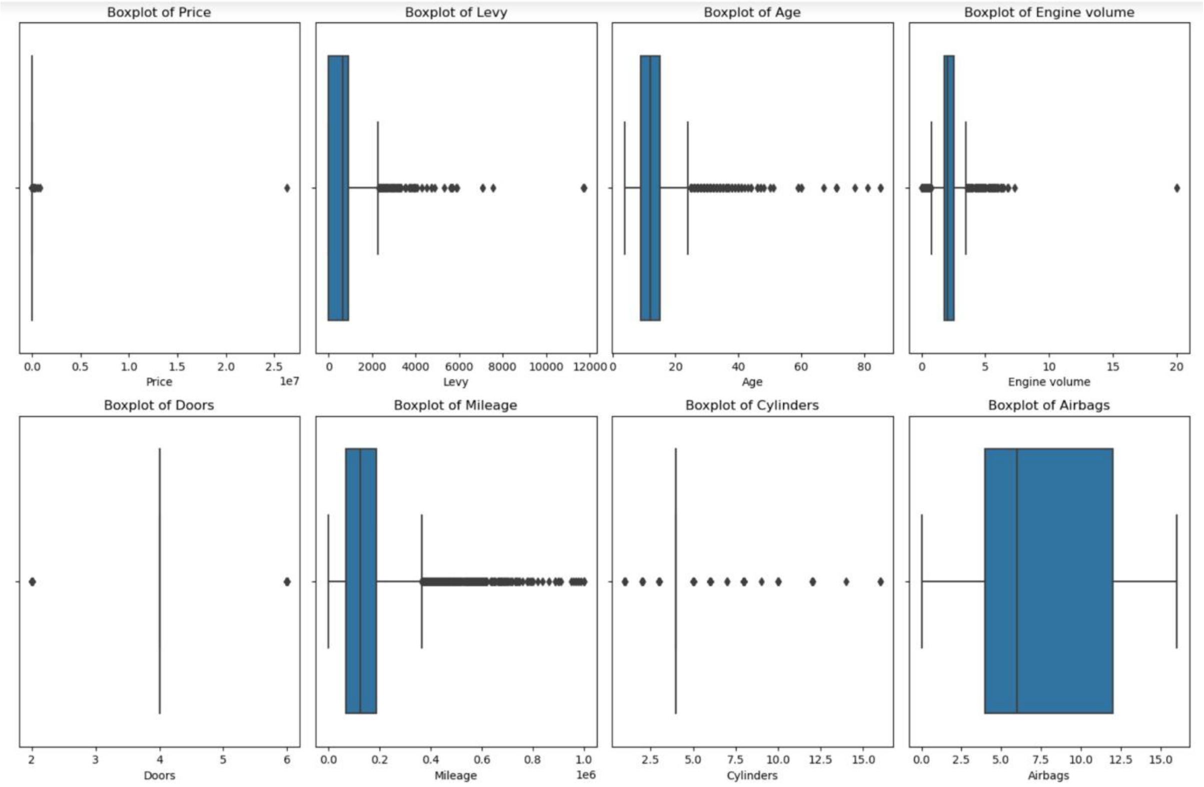


Figure 4: Use boxplot for large-scale outlier handling

Finding that 'Doors','Cylinders', and 'Airbags' have no outliers in the diagram, cylinders are available for all the remaining columns, firstly calculating the first quartile and third quartile of each column and IQR (Interquartile Range) [111], then defining the lower and upper bounds of the outliers. Finally, the value below the lower bound is set as the lower bound, and the value above the upper bound is set as the upper bound. To complete the handling of outliers.

3.1.1.3 Load

Rename and save the converted data, and load it into a Jupyter Notebook for later design.

3.1.2 EDA (Exploratory Data Analysis)

EDA (Exploratory Data Analysis)[14] is a preliminary process of data analysis. It uses statistical charts and visualization methods to discover the structure, anomaly, pattern and correlation of data, which is particularly important before determining the correct data processing and analysis model. The goal of EDA is to use intuitive methods to understand data, which is often a necessary step before modeling.

First of all, this study conducted a visual study on the top ten automobile companies with the largest number of cars sold in the data, and visualized their sales volume and the average price of cars sold, respectively, in the form of bar charts and line charts (as shown below).

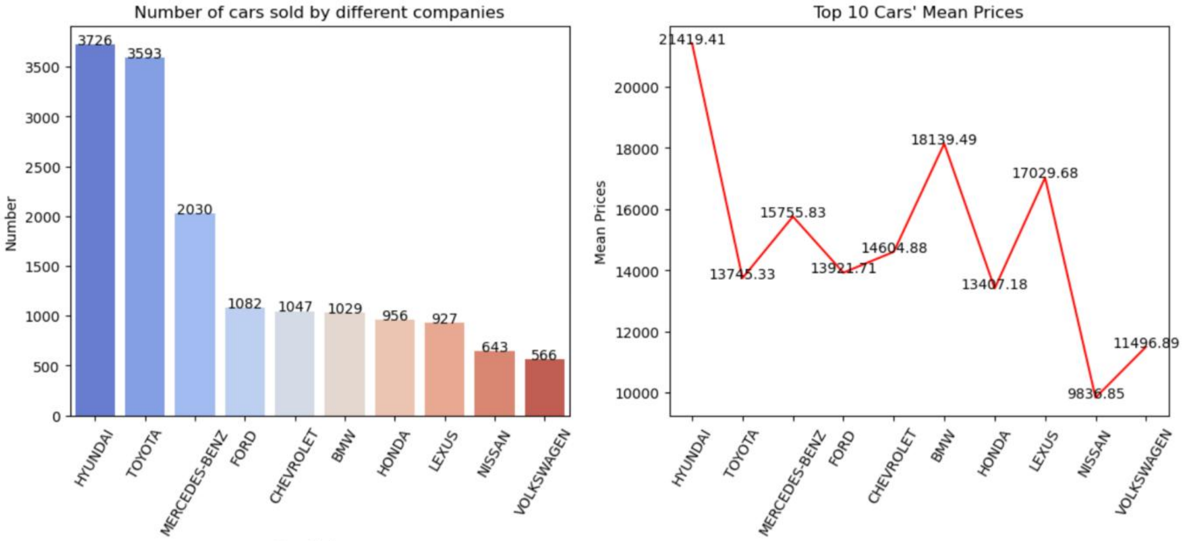


Figure 5: The relationship between sales and prices of the top ten car companies that sell the most cars

As can be seen from the two charts, although some brands have higher car sales, this does not necessarily mean that their average price is also high, and does not show a strong correlation. For example, while Hyundai has the highest sales volume, its average price is in the middle range. Toyota, though slightly behind Hyundai in terms of sales, has the highest average price. That could indicate that Toyota sells more high-end models or that its models are priced higher. Vw, by contrast, has the lowest average price, which could mean it sells more economy cars.

In order to explore the relationship between vehicle mileage and price, this study carried out a visual study, counted the mileage of all vehicles with bar charts, and drew the relationship between the mileage and car price with scattered dots.

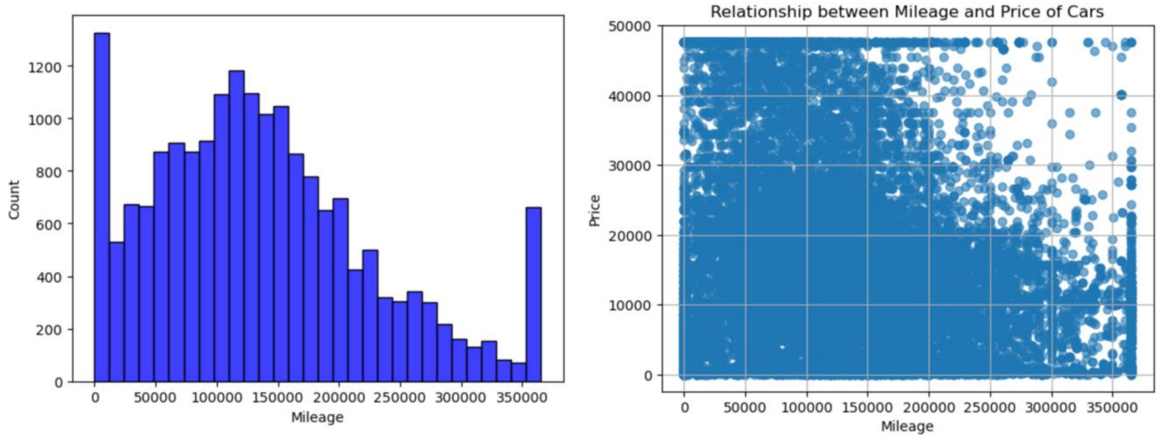


Figure 6: the relationship between the number of miles and the price of a car

As can be seen from the chart, the two variables show a certain degree of negative correlation, which means that the higher the mileage of the car, the lower the price. This is in line with common expectations, as the value of a car typically declines as the number of miles used increases. However, the distribution of data points is very dense, especially in areas with low mileage, which may indicate a very wide range of new or near-new car prices. The range of price changes appears to be more concentrated in high-mileage cars, which may be due to the fact that high-mileage cars are closer in market value, or that consumers buying such vehicles are more focused on price factors.

When analyzing the color variable, the project uses scatterplot to visually explore the data (figure below)

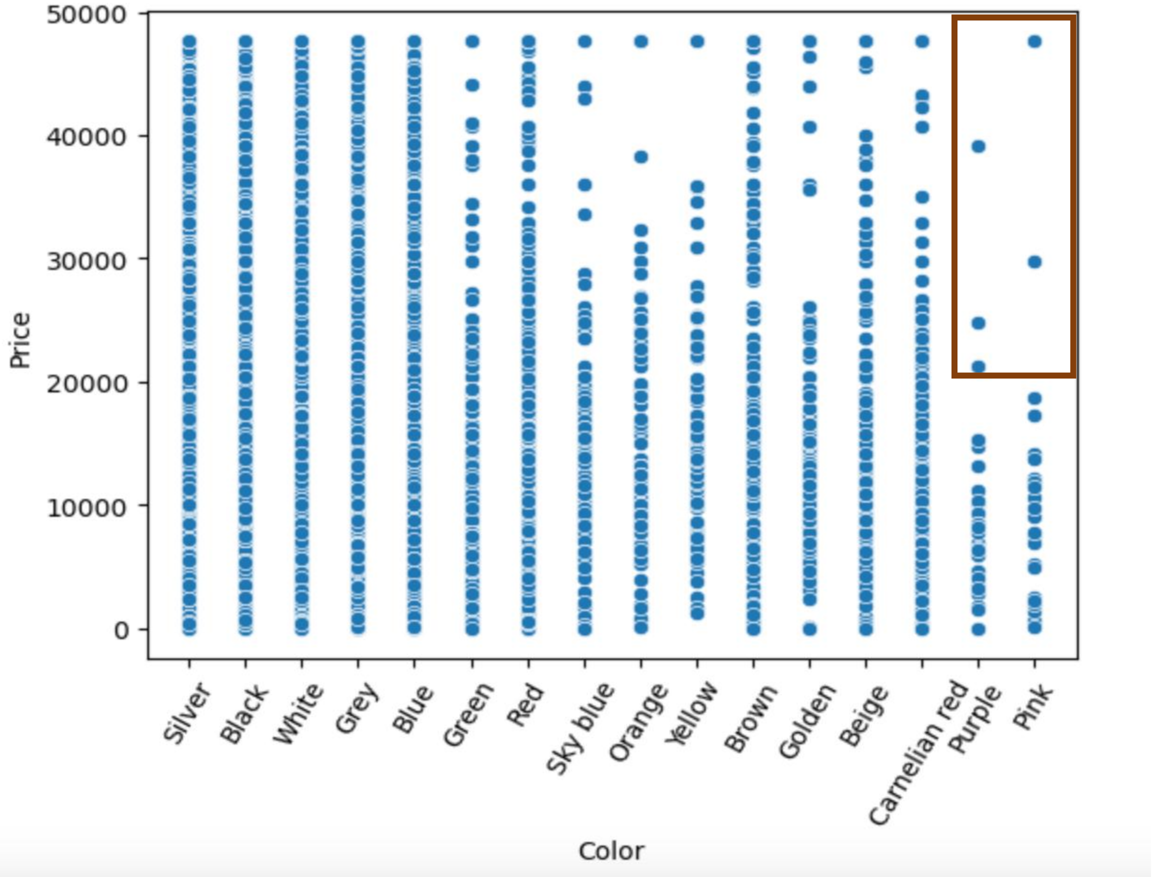


Figure 7: Visualize the column color with scatterplot

It can be found from the figure that the data value in the column color is pink, and the price corresponding to the data in purple is mainly below 20,000. Therefore, this experiment decides to delete the values of pink and purple that are greater than 20,000 (the black border in the figure has been marked), which can make the machine learning results more accurate.

Correlation analysis

This experiment uses heatmap to conduct correlation analysis for each variable that affects car prices (as shown in the figure below).

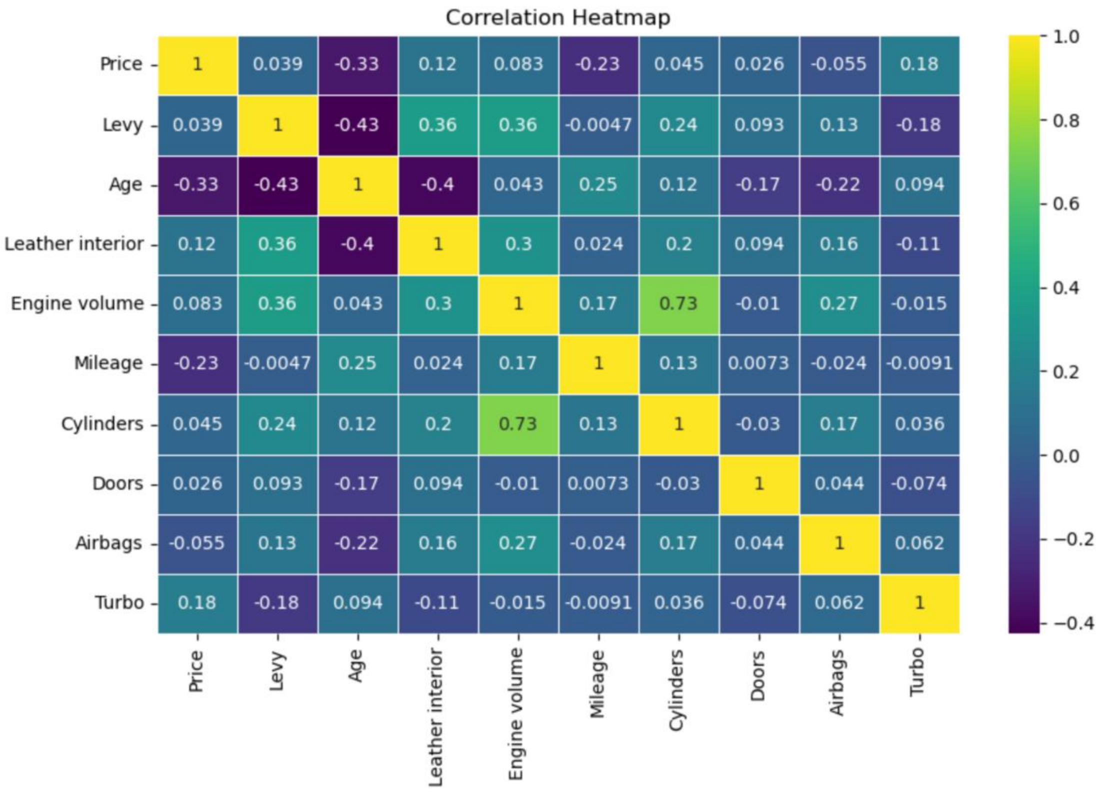


Figure 8: heatmap was used for correlation analysis of each variable

If the absolute value of the correlation coefficient is above 0.5, this study believe that they have strong correlation. If they are between 0.3 and 0.5, this study consider them to have a moderate correlation; If they are below 0.3, this study consider them to have a weak correlation.

Most of the variables in the chart do not seem to have a particularly strong correlation with prices. This does not mean that these variables have no effect on price, because the correlation coefficient of a single variable does not capture the possible complex interactions between the variables, and there may also be non-linear relationships, which cannot be expressed by pure correlation analysis.

3.1.3 Modeling and analysis

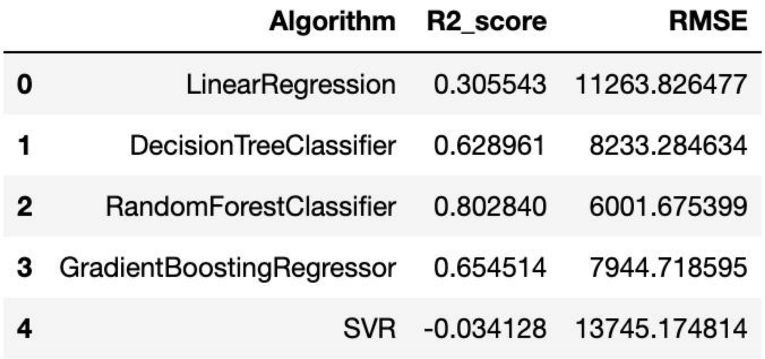
First, all object columns are converted to numeric values by labelencoder, because object columns are not understood by the machine learning model.

Feature selection

test_size=0.25: The size of the test set accounts for 25% of the total data set;

Random_state =5: This is the seed value of a pseudorandom number generator. This value ensures repeatability of results when data is split. This means that any time someone else is running with the same data set and this random seed value.

This project adopts LinearRegression, decisiontreecclassifier RandomForestClassifier, GradientBoostingRegressor, SVR this several models. The initial R2_score and RMSE values for these different models are shown below.



	Algorithm	R2_score	RMSE
0	LinearRegression	0.305543	11263.826477
1	DecisionTreeClassifier	0.628961	8233.284634
2	RandomForestClassifier	0.802840	6001.675399
3	GradientBoostingRegressor	0.654514	7944.718595
4	SVR	-0.034128	13745.174814

Figure 9: R2_score and RMSE value

Grid Search

Grid Search is a method for model hyperparameter tuning by systematically traversing combinations of multiple hyperparameters to find the optimal model setting. This method is widely used in machine learning and is particularly suitable for improving the performance and accuracy of models.[14]

Grid testing works by specifying a parameter grid. The grid contains all the parameters of interest and their possible values.

1. Define a parameter grid: First, you need to define a list of one or more hyperparameters and the range of possible values for each hyperparameter.

2. System Search: Grid testing systematically creates every possible combination of parameters.

3. Model training and evaluation: For each set of parameters, grid testing will train a new model and evaluate the performance of the model by cross-validation and other methods.

4. Choose the best combination: Finally, the grid test selects the combination of parameters that optimizes the evaluation metrics (e.g. accuracy, F1 score, etc.).

Therefore, the project is decided to use the grid search for parameter tuning [15], the project adopted in initial model to show the best RandomForestClassifier and GradientBoostingRegressor parameter tuning.

Result of RandomForestClassifier:

Best parameters: {'max_depth': 30, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 200}

The Score of Model is :0.8079287440236537

Result of GradientBoostingRegressor:

Best parameters: {'learning_rate': 0.1, 'max_depth': 5, 'max_features': None, 'min_samples_leaf': 2, 'min_samples_split': 3, 'n_estimators': 300, 'subsample': 0.8}

The Score of Model is :0.7842764810861698

3.2 Results

In this project, the accuracy of the best random forest model reached 80.79% after the final optimization, which was about 2% higher than the existing experimental results.

A random forest consists of multiple decision trees that can be trained completely independently, making it ideal for parallel processing. Therefore, with sufficient computational resources, the training process of random forests can be significantly accelerated. Random forests are also inherently robust, with less overfitting occurring, especially when there are a large number of trees. This allows random forests to perform well on multiple data sets, reducing the need for tuning parameters.

Because the experimental data in this study are not too strongly correlated, different algorithms may have different performances. The reason why random forest algorithm can perform well in this study is that,

1. Better generalization ability:

By building multiple trees and averaging or majority voting on their results, random forests effectively reduce the variance of the model and improve the generalization ability. This mechanism enables the random forest to be more stable for datasets with low correlation features because it is less dependent on the strength of the association for any single feature.

2. Anti-noise ability:

Since a random forest only considers a random subset of features during each splitting process, this randomness increases the model's resistance to noise. In data with weak feature associations, this approach can reduce the risk of over-reliance on noise features.

3. Not easy to overfit:

The random forest is less likely to overfit the training data as a single decision tree does because it averages the decision results of multiple trees. Even when the correlation between features is low, the random forest can maintain good performance and will not deviate from the correct prediction because of a few insignificant features.

In summary, the advantages of random forest in parallel processing capability, robustness to outliers, and anti-overfitting make it more efficient than GBR and other algorithms in this case.

Chapter 4 Discussion

4.1 Conclusions

This project successfully developed a used car price prediction model based on random forest algorithm. The experimental results show that the accuracy of the model reaches 80.79%, which is significantly higher than the accuracy of about 78% mentioned in the existing literature. This result proves that our efforts in feature engineering, model selection and parameter tuning are effective. The model shows excellent prediction ability in all performance evaluation indexes, especially when dealing with large-scale data sets and complex features, random forest shows its robustness. However, this study also note that there is still room for improvement in the model's predictions in some extreme cases. Overall, this project provides a reliable pricing aid for the used car market and a solid foundation for subsequent research.

4.2 Ideas for future work

Although the performance of the current model is satisfactory, future work can further improve the performance and scope of application of the model in several directions:

Data richness: Introduce more data sources or add more feature variables to improve the model's capture of price sensitivity.

Time series analysis: Considering the influence of market trends and seasonal factors, time series analysis is integrated into the model to predict the trend of price change over time.

Deployment optimization: If the model is more than 85% accurate, the model can be deployed to the cloud to provide users with faster prediction services.

Real-time update: Implement real-time data update and model retraining mechanism, because the current social automobile market is growing and the price fluctuation is obvious, and there are many influencing factors, so as to maintain the timeliness and accuracy of the model prediction.

List of References

Admin 2024. How to Buy a Used Car – Help & Advice. AutoTrader. [Online]. [Accessed 29 April 2024]. Available from: https://shopusedcars.org/?tm=tt&ap=gads&aaid=adaHOrDAEPMRd&gad_source=1&gclid=Cj0KCQjwir2xBhC_ARIsAMTXk84csUGy7acpP7aHJxxu46bU4AHjE_wJr9No8gIDG-j5oqq08W_daIIaArRZEALw_wcB.

Analysis, Z. 2021. Python heatmap. Zhihu. [Online]. [Accessed 25 April 2024]. Available from: https://zhuanlan.zhihu.com/p/364624304.

AndQVQ 2024. ETL (Data Extraction, Conversion, Loading) _etl Data Extraction -CSDN blog. Csdn. [Online]. [Accessed 25 April 2024]. Available from: https://blog.csdn.net/qq_40776361/article/details/122212320.

Anon n.d. Example gallery — seaborn 0.13.2 documentation. [Accessed 25 April 2024a]. Available from: https://seaborn.pydata.org/examples/index.html.

Anon n.d. Examples — Matplotlib 3.8.4 documentation. [Accessed 25 April 2024b]. Available from: https://matplotlib.org/stable/gallery/index.html.

Anon n.d. Gantt chart _360 encyclopedia. [Accessed 30 April 2024c]. Available from: https://upimg.baike.so.com/doc/1132388-1197920.html.

Anon n.d. Quality assurance in Agile development. *Baidu Wenku*. [Online]. [Accessed 30 April 2024]. Available from: https://wenku.baidu.com/view/9506b10d80d049649b6648d7c1c708a1284a0ace.html?_wkts_=1714491645080&bdQuery=敏捷开发质量保证.

Anon 2024a. New Cars, Used Cars, Car Dealers, Prices & Reviews. Cars.com. [Online]. [Accessed 29 April 2024]. Available from: https://www.cars.com/.

Anon 2024b. sklearn.ensemble.GradientBoostingRegressor. scikit-learn. [Online]. [Accessed 29 April 2024]. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html.

Anon n.d. Tuning the hyper-parameters of an estimator. scikit-learn. [Online]. [Accessed 25 April 2024d]. Available from: https://scikit-learn.org/stable/modules/grid_search.html.

Bwcx 2023. Machine Learning outlier Detection and Processing _ Box Plot Detection -CSDN blog. Csdn. [Online]. [Accessed 25 April 2024]. Available from: https://blog.csdn.net/weixin_44639720/article/details/130176743.

CtrlZ1 2023. Hyperparameter tuning method collation -CSDN blog. Csdn. [Online]. [Accessed 29 April 2024]. Available from: https://blog.csdn.net/qq_41076797/article/details/102941095/.

Eu 2018. What is GDPR, the EU’s new data protection law? GDPR.eu. [Online]. [Accessed 25 April 2024]. Available from: https://gdpr.eu/what-is-gdpr/.

JasonZhangOO 2017. Regression evaluation indicators: mean square error root (RMSE) and R square (R2) _r2 and RMSE-CSDN blog. Csdn. [Online]. [Accessed 29 April 2024]. Available from: https://blog.csdn.net/JasonZhangOO/article/details/77725659.

Laru__ 2018. Root mean square error (RMSE), mean absolute error (MAE), Standard Deviation (Standard Deviation); Mean, Standard Deviation, correlation coefficient, Regression Line and least squares -CSDN blog. Csdn. [Online]. [Accessed 30 April 2024]. Available from: https://blog.csdn.net/Laru__/article/details/80756370.

Ll 2024. [Python] What are R2 scores (coefficient of determination) and calculating R2 scores via scikit-learn’s r2_score function -CSDN blog. Csdn. [Online]. https://blog.csdn.net/u011775793/article/details/135444960.

Madhani, B. 2018. The EU.S.GDPR: Lessons for U.S. policymakers - Disruptive competition project. Project Disco. [Online]. [Accessed 25 April 2024]. Available from: https://www.project-disco.org/privacy/052518the-eus-gdpr-lessons-for-u-s-policymakers/?gad_source=1&gclid=CjwKCAjwoPOwBhAeEiwAJuXRh3TtknrinwR73HFWFbKz1bn92fJuh5WBdJe_ja_hYY3gzj9ewnFPFxoCrIQQAvD_BwE.

Maggieyiyi 2023. Machine Learning - Decision Tree _ Machine Learning Decision Tree -CSDN blog. Csdn. [Online]. [Accessed 29 April 2024]. Available from: https://blog.csdn.net/maggieyiyi/article/details/123774872.

NS 2023. Machine Learning - Linear Regression Article _ Machine Learning Linear Regression -CSDN blog. [Accessed 29 April 2024]. Available from: https://blog.csdn.net/weixin_53024882/article/details/131031713.

QuietNightThought 2023. Grid search technology in Machine Learning, How to apply grid search technology in Auto-sklearn -CSDN blog. Csdn. [Online]. [Accessed 25 April 2024]. Available from: https://blog.csdn.net/shdabai/article/details/131215362.

Sethi, A. 2020. Support Vector Regression In Machine Learning. Analytics Vidhya. [Online]. [Accessed 29 April 2024]. Available from: https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/.

Sjysj 2023. Agile Software Development Practices: Team Collaboration and Communication -CSDN blog. *Csdn*. [Online]. [Accessed 30 April 2024]. Available from: https://blog.csdn.net/universsky2015/article/details/137323370.

W3c n.d. Python machine learning. Grid Search. [Online]. [Accessed 25 April 2024]. Available from: https://www.w3schools.com/python/python_ml_grid_search.asp.

zhyl 2023. Random forest in Machine Learning -CSDN blog. Csdn. [Online]. [Accessed 29 April 2024]. Available from: https://blog.csdn.net/lsb2002/article/details/131650614.

Appendix A
Self-appraisal

A.1 Critical self-evaluation

In this project, I made a rigorous self-evaluation of my own performance. I realized that a lot of time and effort had been invested in data preprocessing and feature engineering, which were key factors in the model's good performance. At the same time, I also found that the model selection and parameter optimization needed to be improved. Although the accuracy of the final model was satisfactory, I may have missed some opportunities to optimize the model performance due to my limited ability in the process. In addition, I realized that in terms of project management, I did not plan and manage my project well, and I need to further improve my time management and communication and coordination skills in order to promote the project more efficiently.

A.2 Personal reflection and lessons learned

After this project, I have a lot of profound personal reflection and harvest. First, I realized that in a machine learning project, the importance of understanding and preparing data far outweighs the complexity of the model itself. I also learned how to stay calm in the face of pressure and deadlines and find creative ways to solve problems. Finally, I realized that continuing to learn and adapt to new technologies is the key to survival and success in today's rapidly evolving field. Every challenge is a learning opportunity, and every failure is a stepping stone.

A.3 Legal, social, ethical and professional issues

A.3.1 Legal issues

When implementing the used car price forecasting project, we must ensure compliance with relevant data protection laws, such as the European General Data Protection Regulation (GDPR) [15]. As we process user data, we need to obtain explicit permission for the use of the data and ensure transparency in informing users about how the data is collected and used. In addition, for data sources, such as data obtained through web scraping, it must be ensured that no copyright or use agreement is violated. In this project, all data were obtained legally and used in accordance with the legal framework to ensure the legality of the project.

A.3.2 Social issues

The social impact of the project is mainly reflected in the provision of more transparent and fair price assessment for buyers and sellers of used cars. My goal is to reduce information asymmetries through technology and help consumers make more informed decisions. At the same time, however, I am aware that the spread of technology may have an impact on employment in the used car industry, such as reducing the need for traditional car price assessors.

A.3.3 Ethical issues

In terms of ethics, the project focuses on the ethical use of data and the fairness of algorithms. I ensure that personal privacy is respected in data collection and processing, and that models do not include biased data that leads to unfair price predictions. In addition, I am committed to avoiding any form of algorithmic discrimination and ensuring that models are fair and impartial to all vehicle types and user groups.

A.3.4 Professional issues

From a professional perspective, the project team adheres to the highest standards of ethics and practice. This includes ensuring that all analysis and reporting is accurate, transparent, and accountable for our work. I continue to pursue professional development when conducting model building and data analysis to ensure that the techniques and methods used are up-to-date and most effective. We promote the values of integrity, respect and excellence for all participants in the program.

Appendix B
External Materials

Data source:

https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge

These sites are the existing modeling and accuracy analysis of this project, and are only for reference before the project starts:

https://www.kaggle.com/code/nadaemad2002/car-price-prediction#Model

https://www.kaggle.com/code/esraameslamsayed/car-prices-analysis-and-linear-models#%F0%9F%93%8C%F0%9F%93%8CData-Cleaning-and-Exploratory

https://www.kaggle.com/code/ahmetcalis/car-price-prediction