CAR RESALE VALUE PREDICTION

TEAM ID:PNT2022TMID34636

TEAM LEADER:HARI SHANKAR RAJ.S

TEAM MEMBER:ANNES BELMIN.S

TEAM MEMBER:AMRITHA JESLIN.J

TEAM MEMBER:ANAND.S

REPORT OVERVIEW

The global used car market size was valued at USD 1.57 trillion in 2021 and is expected to expand at a compound annual growth rate (CAGR) of 6.1% from 2022 to 2030. The used car shipment was recorded at 120.3 million units in 2021. The market has witnessed significant growth in the last few years as the price competitiveness among the new players has been one glowing spot in the used car industry. The inability of customers to buy new cars became one of the reasons for the growing used cars sales volume, which is complemented by the investments made by the industry participants to establish their dealership network in the market. These dealership networks helped market participants to brand and make used car options viable. Further, the role of online sales has become a critical growth factor in the market. Online sites in auto marketplaces have played an essential role in bringing access to consumers with a single touch.

A combination of such developments created a significant upsurge in the demand for used cars. In addition, the factors such as affordability, the availability of used cars, the hike in the need for personal mobility, and the emergence of various online players to organize the market have resulted in the growth of the market growth. For instance, in 2019, E-bay Inc. launched a new eBay Motors application to enhance the used car sale and purchase process online. Until recently, automobile manufacturers and dealers have mainly focused on their new vehicle business with the exclusion of used cars, often viewed as a by product. However, the competition in the market and the threat of new entries have created a great extent of an upsurge in the used car dealership. Moreover, the added quality and reliability of used cars changed the consumer attitude and increased the sales of the used passenger cars. Investing in used car management has become one of the market's requirements characterized by a slimming margin, relentless competition, and demanding consumers. Technological advancements such as the development of the internet and the introduction of hybrid and electric vehicles have changed the buyer position in the market. Moreover, consumers are now knowledgeable about the vehicle, their residual value, quality finance charges, availability, the price applied, and sometimes, the profit margin that the seller makes in a closing deal. This knowledge has changed the dynamics and managed to turn customer intelligence to their advantage. Resultantly, consumers are more inclined toward buying used cars nowadays. Some of the key factors including transparency and symmetry of the information among the consumers and buyers, the online sales channel growth, certified used vehicle programs, and the strong position of franchise dealers play a vital role in driving the market for used cars. Various leading companies have set up online and offline stores worldwide to offer seamless used car buying experiences. For instance, in September 2020, AutoNation Inc. expanded its pre-owned vehicle store and opened two new stores in the USA Denver market.

The organization also announced its goal to open 130 AutoNation Inc. stores in operation across the USA by 2026. In both developed and developing countries, the used to new vehicle ratio has increased in the last few years, accounting for the reasons stated earlier. In addition, franchised dealers with support from OEM involvement in certification and marketing programs, online inventory pooling, and access to high-quality contracts are in a strong position to benefit from the growth in the market. The COVID-19 (Coronavirus Disease) pandemic has placed the automotive industry at great disruption. In the aftermath of the pandemic, the consumers are expected to prefer private conveyance. However, the financial disparities are expected to hamper the purchase of new vehicles and due to budget constraints, commuters are expected to opt for the used cars. Further, virtual reality, online, or digitally generated sales leads buy new vehicles in this pandemic period. Due to the pandemic, hybrid and electric vehicles are expected to battle in the market for the next two or three years due to the current economic conditions.

Research on Second-hand Vehicle Evaluation System Based on Improved Replacement Cost Method

In recent years, China's second-hand car trading volume and transaction amount have increased year by year. The transaction amount is from 315.14 billion to 553.54 billion since 2011 to 2015. In 2015, the transaction volume of used vehicles have reached 9,414,700, so the valuation of used vehicles became a hot spot of concern and research. The improved replacement cost method is widely used in second-hand car value assessment. This article designs a second-hand car value assessment system based on the improved replacement cost method.

Adjustment of purchase price

In order to improve the authenticity of the second-car price assessment, it is necessary to readjust Advances in Engineering Research, volume 163 1291 the evaluation price of the second-car according to the vehicle's own market supply and demand, market competition, routine maintenance, and the technical condition of the vehicle itself. We use the analytic hierarchy process to determine the weights of different factors, and adopt expert scoring to construct judgment matrix.

Design of second-hand car valuation system based on improved replacement cost method

When designing the system, the second-car evaluation process should be analyzed at first. And then the function module and database of the second-hand car evaluation system are designed, and finally the interface design of the system is completed.

Second-hand car identification assessment system hierarchy

In the logic design of the used car value evaluation system, the system's calculation work is dispersed in the application program. The designed system can complete the input of used car information, store the input in data form, and process the data according to the operation instructions of the improved replacement cost method. The system uses B/S structure. Presentation layer: It mainly implements the interaction between users and systems and consists of Web pages with different functions. The user presentation layer displays the system to different users. The user makes requests to the business logic layer through operations such as selection and input. The result of the process is returned to the presentation layer and displayed to the user. Business Logic Layer: This layer is the middle layer between the presentation layer and the data access layer and is the most critical part of the system. This section uses IIS on the Windows platform and programming language uses ASP. This layer composed of upload components and file management components. The upload component can upload the picture information of the used car; the file management component can add, delete, and other operations on the file. Data access layer: It mainly saves and reads data, and reads the information stored in the database to the business logic layer. Data reading is done by the ADO component, and the data processed by the business layer is saved in the database. ADO components include Connection, Command, and Record set. Connection is used to establish a connection with the database ; Command is used to execute commands on the database, such as query, add, delete, modify records, etc. The database used in this design is Microsoft SQL Server 2008.

System Process Analysis

When evaluating a second-hand car, selecting the appropriate valuation model is necessary. According to the need to enter the evaluation parameters, different valuation models need to enter different parameters. According to the evaluation parameters input by the system, the client verifies the validity of the entered form data through the data verification control. If the verification is valid, we should submit the entered form data to the server. The server generates an improved replacement cost method valuation formula to calculate vehicle prices based on formulas and parameters, and save the calculation result to the database, and return the calculation result to the client as well as.

Design of system function module

The second-hand car value assessment system is mainly used to evaluate the value of Advances in Engineering Research, volume 163 1292 second-hand car and provide information on the valuation of used cars. The system mainly includes a system management module, a second-hand car parameter management module, a second-hand car evaluation management module, and an evaluation information query module. System management module is mainly to achieve user management and system login. After the user is registered, the user name and password of the system are obtained. If the user name and password are correct, the user enters the system and check related information. If you make a mistake, relevant tips should be made. The administrator of the system can manage the registered user information, including adding, deleting, etc., but cannot modify the related parameters of the system. The parameter management module is mainly used to input the relevant parameters of the second-hand car, calculate the new rate of the second-hand car according to the corresponding calculation formula, and display the related calculation result. The system administrator can add and retrieve value of the second-hand car's vehicle condition parameters and new rate options. The second-hand car evaluation management module is mainly to obtain the evaluation price of the second-hand car and to evaluate the sale price. Click Save to save the data information in the system database. The evaluation information query module stores vehicle information that has been evaluated. Through this module, the parameters of the second-hand car, the evaluation price, and the evaluation price of the second-hand car can be queried.

Vehicle Type Insights

The conventional vehicle segment accounted for a share of over 40.0%, in terms of shipment, in 2021. The electric vehicle segment is expected to register a significant CAGR over the forecast period, complemented by the hybrid vehicle. In the last few years, used electric vehicle prices continue to remain viable for consumers, and this plays a significant driving factor for electric vehicle sales. According to the last few years' price analysis, used electric vehicles' prices have been lower than the used hybrid vehicles. Electric vehicle traits such as technology-driven performance, in the luxury vehicle segment, provide a status symbol and support sustainability, thus creating a significant volume demand for used EVs. Conventional gasoline vehicles with large inventory offer multiple choices at an affordable price. This segment of vehicles accounted for the maximum share in all sizes, including compact cars, mid-size, and SUV cars. Further, growing concerns over climate change and increasing pollution have created a great demand for a substitute for conventional gasoline vehicle. Hence, there has been significant growth registered by the electric used cars in the market.

Vendor Type Insights

The organized vendor segment accounted for the largest volume share of over 70.0% in 2021. This is attributed to the increasing number of franchised dealers in the market. The entry of new players in the market and new retail models also emerged as a key factor in fueling the growth of the market. According to the NADA, in the U.S., franchised dealers earned higher gross profits on used vehicle sales than independent dealers. In addition, the organized vendors benefited from greater consumer loyalty to the brand across all age groups. The organized vendor segment is expected to witness high growth over the forecast period. The segment is accepted to hold more than two-thirds of the market in the coming years. With many dealers across the globe, the market is highly fragmented. However, in developed countries such as the U.S., Germany, and the U.K., there are some top dealers such as CarMax Business Services, LLC and Asbury Automotive Group that account for more than half of the volume share in the market.

Regional Insights

Asia Pacific accounted for the largest share of over 35.0% in 2021, in terms of shipment, majorly due to the rapid growth of demand in China for used vehicles. Asia Pacific is projected to expand at the highest CAGR over the forecast period. This is attributed to the increasing sales of the used car in China, India, and other Asian countries. The North American region held a notable market share in 2021 and is expected to witness steady growth in the years to come owing to the plummeting growth in the past few years.In the Asia Pacific region, with the rising number of organized players with used car trading services, China has expanded its market footprint in the Asian market. Some Indian car dealers provide a rich array of advanced technology-enabled tools, which include mobile-based applications, a virtual online showroom, cloud services for lead management systems, tracking sales performance, and digital marketing support. Moreover, this extent of advancement in the Indian used car industry creates great opportunities for the consumer base. Within the region, Indonesia, Malaysia, Indonesia, South Korea, and other developing countries have shown significant potential for the market.

Fuel Type Insights

The petrol segment accounted for the largest volume share of over 40.0% share in 2021. This is attributed to the declining usage of diesel vehicles as the government discourages the purchase of used diesel vehicles. The others segment is expected to witness significant growth over the forecast period. In developing countries, CNG powered vehicles have also shown a sustainable upsurge in used vehicle volume sales. Emission standards for the positive ignition (gasoline, NG, LPG, ethanol) and compression ignition (diesel) vehicles have become one of the reasons for the slump in sales of diesel vehicles. Moreover, excessive emission of NOx by the diesel engine can be attributed to the decline in diesel engine vehicle sales and an increase in the substitute market. The petrol-fueled car emission standard is less stringent compared to diesel-fueled passenger cars. Furthermore, petrol cars with a refined engine, decent fuel efficiency, and strong top-end performance attracted a large consumer base in the last few years and are expected to continue with the same in the coming years. In addition, increasing inventory for petrol-based SUVs became one of the driving factors of the petrol segment.

Size Insights

The SUV size segment accounted for the largest volume share of over 35.0% in 2021. With the changing landscape in the automotive market, the SUVs segment has caused the downfall of other segments. Offering space and size while remaining compact compared to off-road vehicles, SUVs are considered ideal drives by buyers nowadays in various regions. With great demand and a wider supply network, residual value for SUVs is higher nowadays for the market. The European region has witnessed significant demand traction for the used SUVs market. The compact size segment is expected to register a significant CAGR over the forecast period. This is attributed to people's preference for economical and compact size vehicles. Compact size vehicles with a high production rate and huge inventory have been preferred among the franchised owners. Easy availability with affordable prices fuelled the demand for the used compact vehicle in the last few years. However, with the changing consumer preferences and advancements in SUVs, the used SUVs have shown significant growth and this is expected to continue in the coming years.

Key Companies & Market Share Insights

The key players in the market are focusing on expanding the customer base to gain a competitive edge in the market. Thus, vendors are taking several strategic initiatives, such as collaborations, acquisitions & mergers, and partnerships. For instance, in 2020, Volkswagen announced a major investment in the market for used cars by a collaboration of its own used-car chain, Das Welt Auto, with various used car platforms. Mainstream automakers have also been expanding their presence in this space with their pre-owned car sales networks like Maruti Suzuki's True Value, M&M Mahindra's First Choice Wheels, and Toyota's U Trust. Some prominent players in the global used car market include:

1. Alibaba.com
2. Asbury Automotive Group
3. AutoNation Inc.
4. CarMax Business Services, LLC
5. Cox Automotive
6. eBay Inc.
7. Group 1 Automotive Inc.
8. Hendrick Automotive Group
9. LITHIA Motor Inc.
10. Scout24 AG
11. TrueCar, Inc.

Segments Covered in the Report

This report provides forecasts for revenue and volume growth at the global, regional, and country levels and provides an analysis of the latest industry trends and opportunities in each of the sub-segments from 2017 to 2030. For this study, Grand View Research has segmented the global used car market report based on vehicle type, vendor type, fuel type, size, sales channel, and region:

- **Vehicle Type Outlook (Shipment, Million Units; Revenue, USD Billion, 2017 - 2030)**
    - Hybrid
    - Conventional
    - Electric
- **Vendor Type Outlook (Shipment, Million Units; Revenue, USD Billion, 2017 - 2030)**
    - Organized
    - Unorganized
- **Fuel Type Outlook (Shipment, Million Units; Revenue, USD Billion, 2017 - 2030)**
    - Petrol
    - Diesel
    - Others
- **Size Outlook (Shipment, Million Units; Revenue, USD Billion, 2017 - 2030)**
    - Compact
    - Mid-size
    - SUVs
- **Sales Channel Outlook (Shipment, Million Units; Revenue, USD Billion, 2017 - 2030)**
    - Online
    - Offline
    -

Data Sets

Kaggle Dataset The dataset was sourced from Kaggle and includes 122,144 car listings from the years 2018, 2019, and 2020 from all areas in the United States. It is available publicly. It includes all types of road-going consumer Figure 1: UML Component Diagram, Implementation Overview 22 vehicles, such as vans, pickup-trucks, and cars (See Appendix B for the published data set). This dataset shows listings of used cars and not necessarily the final sales price. The dataset does not have duplicate listings of the same car however, with the previous listings being removed as the sale was most likely unsuccessful. Therefore, the listed price may be somewhat higher than actual sales price and not reflective of the actual values of the cars. This error is consistent across the dataset (including the entries that will be used for testing) however and should not significantly affect measured model performance.

Data Cleaning and Normalization

The first step in cleaning the dataset provided from Kaggle was to identify variables which will not be useful for training the models. This includes features which are not correlated with price, have too many discrete values to draw inferences from, or have too many missing values. The features that were identified to be dropped from the dataset were: ID, zip code, and Trim. The next step is identifying and removing outliers for the ten remaining features. Keeping in mind the distribution of the data and the negative effect of removing too many values, appropriate minimum and maximum values were set for each feature to remove rows in the dataset which were extreme in any feature category. This was performed for the features prices old, Mileage, and Year. (See Table 2) These were chosen somewhat arbitrarily but with the purpose of removing an appropriate percentage of uncommonly occurring extreme values in the dataset. This increases the performance of the models.

Machine Learning Algorithms

The data, after being cleaned and normalized, is split into training and test data using a randomized 80-20 split. This is to ensure that the data used for testing does not contain any of the data used for training. Thus 20% of the data is reserved for testing purposes (see 4.4 Inference). The training dataset was used to train the four price prediction ML models chosen: Multiple Linear Regression, Lasso Regression, Ridge Regression, and Random Forest Regression. All machine learning algorithms used in this report were imported from the s k learn library. Some models were provided input parameters to implement. The motivations for the choice of input parameters are explained in this section for the models that require them. The Ridge Regression model was implemented with the argument alpha=0.01. An assortment of different alpha values were tried, and lower values performed slightly better. Values lower than 0.01 didn’t noticeably perform improve model accuracy. Lasso Regression was similarly implemented with an alpha=0.01 value after testing. Lower values gave better prediction results. Random Forest was implemented with default parameters and random state=0. The random state is necessary because it is a stochastic process that takes a seed value to begin. When testing different values, there was no noticeable performance increase. The random state throughout training was therefore set consistently to 0, minimizing stochastic behaviour resulting from varying the random state.

Inference

Inference involves using the subset of the data that was reserved for testing (20%) to predict the price based on the features. This step was performed after the dataset was cleaned and normalized, and the models were optimized. The dataset was re-split, models were retrained, and inferences retaken a total of five times. This produced five separate inferences with the same parameters to be able to produce an average for the measurements. The inferences produced varies slightly each time as a result of the randomized 80-20 training-testing data split. Each model produced inferences from the same testing subset in every iteration. To judge overfitting, they were also tested on the training subset of the data. 26 Much better prediction results on the training data is an indication of overfitting.

Measurements

The measurements taken for this study are described in this section. All of the measurements are taken from the same inference data for each model and using the formulas for the various metrics.

Training and Testing Accuracy Comparison

The three performance metrics that were taken for machine learning algorithm are R-squared, RMSE, and MAPE. These measurements were taken for both the training and testing inferences and averaged across all five iterations of inferences taken to produce a table of metrics. The MAPE is of special importance for evaluating the potential of the algorithms to be used for a consumer valuation tool and fulfilling Research Question 3. Therefore, the dataset was split into four price categories and MAPE measurements taken for the inferences each algorithm. The four price categories were each approximately 25% of the dataset each, with the first one being the lowest priced cars and the last being the highest priced cars. In other words, the MAPE was taken for cars belonging to the 0th-25th price percentile, 25th-50th price percentile, 50th-75th price percentile, and 75th-100th price percentile. This serves to demonstrate the performance of each algorithm across different price categories of cars.

Inferred Price Plots

For the first iteration of inferences (both training and testing), scatter plots were created for each algorithm where each data point is the actual price, plotted against the inferred price. A line of best was then calculated and drawn through these points as well as a line showing the actual values, y=x. If the algorithms do not demonstrate any systematic error, the line of best fit should match this line. See Appendix A for these scatter plots.

Analysis and Discussion of Results

Training and Testing Accuracy

The measurements show that the RSME value for the testing dataset of Linear Regression was the same as for Ridge Regression at 5953, while Lasso Regression performed slightly better at 5950. Random Forest had a much lower value at 4799. Linear, Ridge, and Lasso had a very similar RSME for training data that was somewhat lower than their respective RSME's on the testing dataset. This is to be expected, since the models are trained to minimize the squared error on the data it is trained on. The Random Forest Regression algorithm had an RSME of 1975 for the training data. The R-squared error for each similarly showed that Random Forest Regression had much better performance. The rest of the algorithms performed slightly worse on the testing data than the training data. Random forest had a value very close to 1, which is the highest possible value that is only reached when each price in the data set is predicted perfectly. This indicates overfitting. This occurs when the model is too complex for the data, and over-tunes the coefficients to predict the individual data points in the training set, while not generalizing well for unseen data (testing data). Linear Regression, Ridge Regression, and Lasso Regression achieved similar performance to each other judged by the MAPE metric. Interestingly, they all demonstrated better performance on the testing data than the training data for one of the five iterations that were averaged to produce the table. The algorithms are all defined in the sk learn library to minimize the RSME on the training data rather than MAPE. Therefore, the algorithm is not optimized to give the lowest possible MAPE value for the training data, although this value is related to the RSME. The testing data is 25% the size of the training dataset and will therefore give more variance in the results. By chance the testing data 33 can achieve a lower MAPE than the training data and did so in one of the five iterations. The Random Forest Regression MAPE value for testing data shows that it performed better in its overall score, with a value of 37.65% on the testing data compared to the next best score of 44.45%. When examining this value by four different price percentiles, Random Forest showed higher performance in all four categories although the 0th -25th and 75th - 100th percentile categories showed the largest increase in performance relative to the other algorithms (see figure 6). The histogram depicting the density of model predictions (see figure 7) shows that the distribution of the predictions are, for all of the models, very concentrated at roughly 5,000. Appendix A shows plots of the predicted values against the actual values for each algorithm. The line of best fit shows the center point of the line and demonstrates that Linear Regression, Ridge Regression, and Lasso Regression plot have a systematic error in the concentration of predictions for higher actual car prices. The algorithms are very likely to predict the price as lower than actual.

Magnitude of Coefficients

The magnitude of the coefficients for the Linear Regression model were very large as shown in Table 4. The coefficients of the Ridge Regression and Lasso Regression algorithms, which both penalize the use of large coefficients, were much smaller in comparison. This did not seem to have a significant impact on any of the performance metrics, as these three algorithms performed very similarly in all of them. A potential implementation-dependent issue is rounding errors stemming from the use of very large magnitude metrics which approach the upper and lower bounds of values of the datatype that is used to store the coefficients. Since no significant differences in performance was shown in the measurements, there didn't seem to be any such issues.

Dataset Limitations

A limitation on any ML algorithm's performance is the dataset. If the dataset does not include features that are strongly correlated to the price, the ML algorithm might not have access to enough information to accurately infer the price. Some strongly correlated features can be rendered redundant if another feature is included in the dataset and is strongly correlated to the redundant feature. If this is the case for a missing feature, it may be partially redundant and therefore unnecessary 34 to include in the dataset. The dataset that was chosen for the training of the models in this work initially included a feature for the zip code for the sale. This feature was removed as part of the Data Cleaning and Normalization outlined in the method. A previous report by Sri Totakura and Harika Kosuru [12] comparing ML Regression model performance found that the "Region" feature in their dataset had the highest feature importance. The study concluded that this feature had the highest correlation to the price by comparing each feature's impact on the price compared to the rest, for their best performing algorithm (Light Gradient Boosted Machine). This suggests that deriving a "Region" feature of the car sales from the zip code available in the dataset could improve model performance. As it relates to the research questions to be answered in this work, the increase in performance could differ between models and affect the results of this study for comparing the performance of models.

Measurement of Average Simulated Depreciation

The average depreciation was shown to be the same for Linear Regression, Ridge Regression, and Lasso Regression, independent of the age of the vehicle. This value was approximately 9.7%. The Random Forest Regression algorithm showed approximately 13.2% depreciation for cars that were 2-4 years old, and approximately 14.6% for cars that were 13-15 years old.

Project method discussion

As described in the project method description, this work utilizes quantitative analysis with measurements of the implemented machine learning algorithms to achieve the purpose of this work and answer the research questions. To be able to take the performance measurements required as part of this, the machine learning algorithms to compare had to be implemented correctly and using a suitable dataset. The requirements for this dataset were for it to be large, be representative of the used car market which a consumer valuation tool is likely to target, span several years in its sales, and include many relevant features.

The project milestones were chosen to accomplish this. The project milestones for this research were all met. As part of Milestone 1, four machine learning algorithms were identified as the best candidates for comparison, based on previous research in similar areas 35 and which are commonly utilized for regression analysis. Milestone 2 was to choose an appropriate dataset. The dataset chosen met the requirements, although it would potentially have increased the performance of the ML algorithms if it had exceeded the requirements to a greater degree.

Because of missing information in some of the features, most of the cars in the dataset were dropped as part of the data cleaning process. Additionally, more features that could be retained after the data cleaning process would be beneficial. Particularly the “zip code” feature, which gives information on the region had too many discrete values to be made continuous. A dataset with more localized sales and therefore fewer discrete values would be more useful. It is possible that the dataset is not representative of the used car market, that is to say that cars with certain features are more likely to be included in this dataset and overrepresented. Milestone 3 was to make appropriate normalizations to the data. This was accomplished but at the cost of reducing the size of the dataset and number of features. Milestone 4 was to implement each of the models chosen as part of Milestone 1. The data cleaning and normalization process was refined based on the requirement for the implementations. The same data was used for all of the models, and the data cleaning and normalizations made should be made to suit all of the models, not to increase the performance of any one model. This was to ensure fair comparison in the performance measurements. Milestone 5 was to make the performance measurements. The metrics for this were chosen to best answer the research questions. To evaluate performance, two metrics commonly used in ML and related to the loss function that each model optimizes during training were used. Additionally, emphasis was put on using a third metric that is more useful for evaluating use for a consumer tool, for several price categories of cars. Milestone 6 was to compare the predicted depreciation for each model. This was to answer research question 2. There are several alternative approaches to doing this, but the one used in this research was to average the percentage decline in predicted price for all cars of a certain year in the testing dataset, and for different ages to see how well the models predict geometric depreciation. Since it is unlikely that the same car was sold multiple times in the dataset, the true depreciation could not be known to evaluate the performance. Instead, the measured average depreciation was compared to an average for all cars obtained from previous research in this area. For this comparison to be valid, the dataset needs to contain years where the depreciation followed this typical average depreciation, 36 and the cars included in the dataset need to be representative of the total market. If these assumptions are true, then any deviation from the average can only be explained by inability of the models to detect depreciation.

Scientific discussion

The results of this study show that Random Forest Regression was able to achieve better performance for the prediction of price for used cars than the other algorithms tested. Further, the three others tested are variations of Linear Regression, and all performed very similarly despite large differences in the magnitude of coefficients. Random Forest Regression scored better on all three commonly used ML regression metrics and assessed depreciation much more accurately. This makes it more suited to developing a consumer tool for price prediction. However, even when broken down by price category, the model did not achieve a lower MAPE than 20%. Previous research into housing was able to achieve a MAPE of 6.37% for housing price prediction. The conclusions drawn are also limited by the weaknesses of the dataset and model implementation. The dataset was filtered to exclude cars with very low or high values in any feature category, as well as rare car brands. For cars with these excluded values, the model may not be able to predict their prices well and therefore conclusions for the performance of the models may not be applicable to them.

Ethical and societal discussion

This work utilized a public dataset published on Kaggle (see Appendix B). This dataset was web scraped and this web scraping must be handled ethically. Webscraping is the use of automated tools for collection and extraction of data from the Web for use of further analysis of this data. Web-crawling is one of these techniques that involves running a script that automatically browses a website and retrieves data. This was done by the creators of the dataset, who web-crawled Ebay.com to gather the data. The legality of web scraping depends on the terms of use of the website, infringement of copyright for commercial use, and if any damage occurred to the website. In addition, there are ethical concerns to consider. The privacy of data for individual users of the website must not be compromised. [13] 37 This research is intended to expand the knowledge needed to create of consumer tool for valuation of used cars using Machine Learning. Such a tool has the potential to change the market for used cars. The societal impact of this tool for consumers looking to buy and sell cars could, if handled responsibly, increase visibility and equality in the market for used cars, as far more individual factors could be considered for valuing a used car. This tool in the hands of an un-informed buyer, could ensure that they are receiving a fair price, and bypass the need for trusted "middle-men" to facilitate a sale. The elimination of "middle-men" for a transaction means lower frictional costs in the market, and a potential for the seller to find a buyer more quickly and for a higher price. If the use of this tool is widespread and consumers base their buying and selling decision on its predictions, the price of cars could be influenced by the predictions. Errors and other problems with the predictions or tool could negatively affect some consumers. In addition to the impact for individual consumers, the car market and industry that facilitates the sale of used cars could change fundamentally. This could mean the loss of jobs or other negative effects.

Conclusions

This article has designed the used car value evaluation system, which is based on the improved replacement cost method. The results show that the system application effect is better. Due to the short system design time, the system's function is not perfect. In the next step, the actual application effect of the system will be tested and evaluated. It is expected that the system will eventually meet the automotive evaluation requirements. In this paper, we have trained our model with used cars data set to predict the price. Here we have used the K nearest Neighbour algorithm and we got accuracy 85% where the accuracy of linear regression is 71%. The proposed model is also validated with 5 and 10 folds by using K Fold Method. The experimental analysis shows that the proposed model is fitted as the optimized model. In our future work, we will apply advanced machine learning techniques and validate the model with different methods to enhance the optimization of the model with improved accuracy.

References

[1] Tu Weixing. Research on Used Car Evaluation System[J]. Nanjing: Nanjing Forestry University, 2008.

[2] Liu Lijuan，Liang Tianran. Replacement Cost Method——Discussion on the Appraisal Method of Used Car Identification [J]. Chinese business community, 2010(9): 37-38

[3] Zhou You. Research on the Valuation Method of Used Car Based on Replacement Cost Method [D]. Jinzhou: Liaoning Agricultural University, 2014.

[4] ZHU Si-hong, ZHANG Chao.Study on PID Control Simulation of Tractor Electro-hydraulic Suspension System [J]. China Manufacturing Information, 2008,37 (21): 49-53. [5] Wei Wei, He Yan. Intelligent control basis [M]. Beijing: Tsinghua University Press, 2008: 101-166