Denise Dodd

Used Car Prices

White Paper

Business Problem

Buying a used car is an opaque process for the average buyer. Without an understanding of the proper sale price of a used car and the factors that impact the price, a buyer will not be able to properly budget and narrow down their car search. Researching cars that are above a buyer's budget can be a waste of time for both the buyer and the seller. Testing cars that are below a buyer's budget can result in the purchase of a car that does not include all the buyer's desired features. In this study, I will use common features of used cars to train a predictive model to provide an estimated sale price. I will also use correlations and important features to determine how each variable affects the price of a used car. At the end of the study, a user will be able to input the features of their desired used car and obtain an estimated sale price. They will then be able to modify their desired features to adjust the anticipated sale price until it falls within their budget.

Background/History

For this study, I will be utilizing a dataset titled "Car Price" (Brar, 2024). This is a dataset with details of over 8,000 used cars. Variables include numerical variables of varying units of measurement including selling price, kilometers driven, engine power, number of seats and the production year of the car. It also includes categorical variables such as the type of fuel the car takes, what kind of transmission that car has, what the prior ownership status of the car was, and if the car is being sold via a dealer or an individual.

Data Explanation (Data Prep/Data Dictionary/etc)

Before I can prepare the data for my pipeline and regression models, I must clean and validate the data set. In the process of cleaning the data, I changed the data types of variables, created and deleted columns, visually assessed the spread of the data, analyzed outliers, resolved entries with null values, and renamed entries for consistency in the data set.

Once the data was cleaned, I separated the columns into features and target variables (with selling_price designated as the target) and further divided the features and target variables into 80/20 training/testing sets. I prepared the data for my pipeline by assigning the numerical features columns to one variable and the categorical features columns to a second variable. I used pipelines to scale the numeric variables (due to the varying units of measurement) and make dummies (via One Hot Encoding) of the categorical variables.

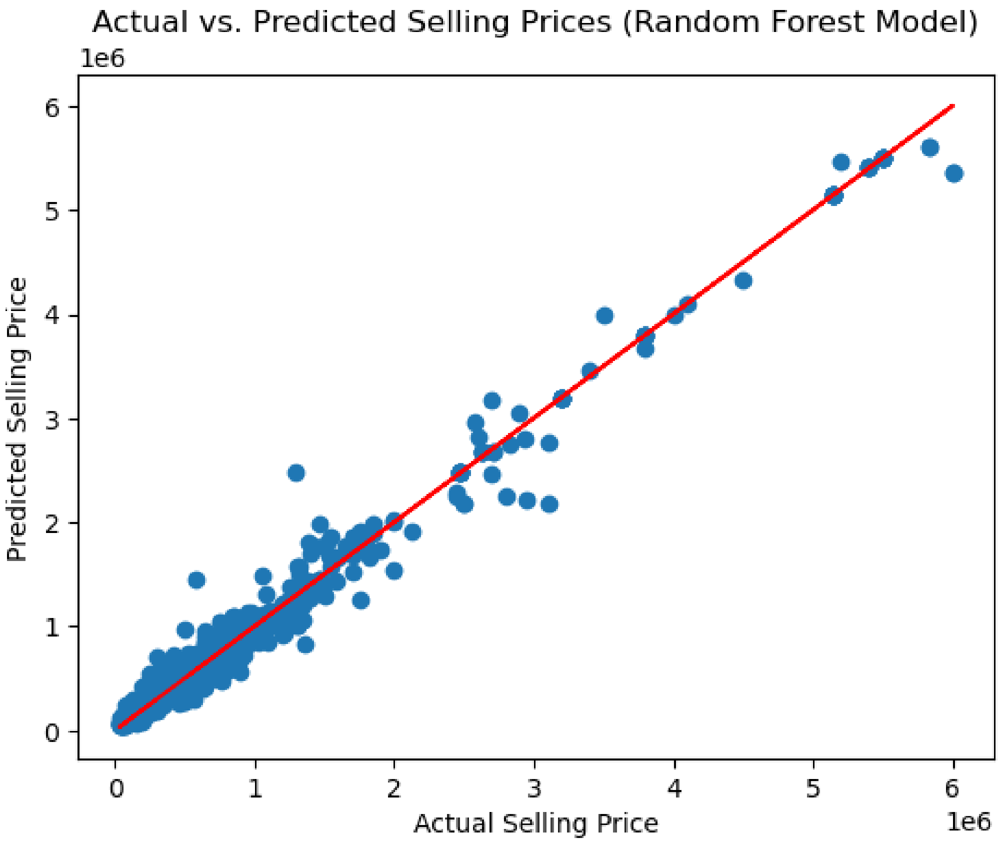
Methods/ Analysis

With the data prepared, I ran several regression models through a loop which used the data transformed via the pipeline to fit the models, make predictions, and calculate evaluation metrics. Both Random Forest and Decision Tree models had similar metrics, but Random Forest proved to be slightly more efficient. This is not surprising as the Random Forest model is an ensemble of several Decision Tree models. The Random Forest model had an R-Squared value of .9834 meaning that 98.34% of the variance in selling price (target variable) can be explained by the model. This model also showed the lowest RMSE with a value of 102,669.73 indicating that on average our predictions will be off by this amount. The range of the target variable is from 29,999 to 10,000,000 so an average outage of 102,669.73 is not concerning.

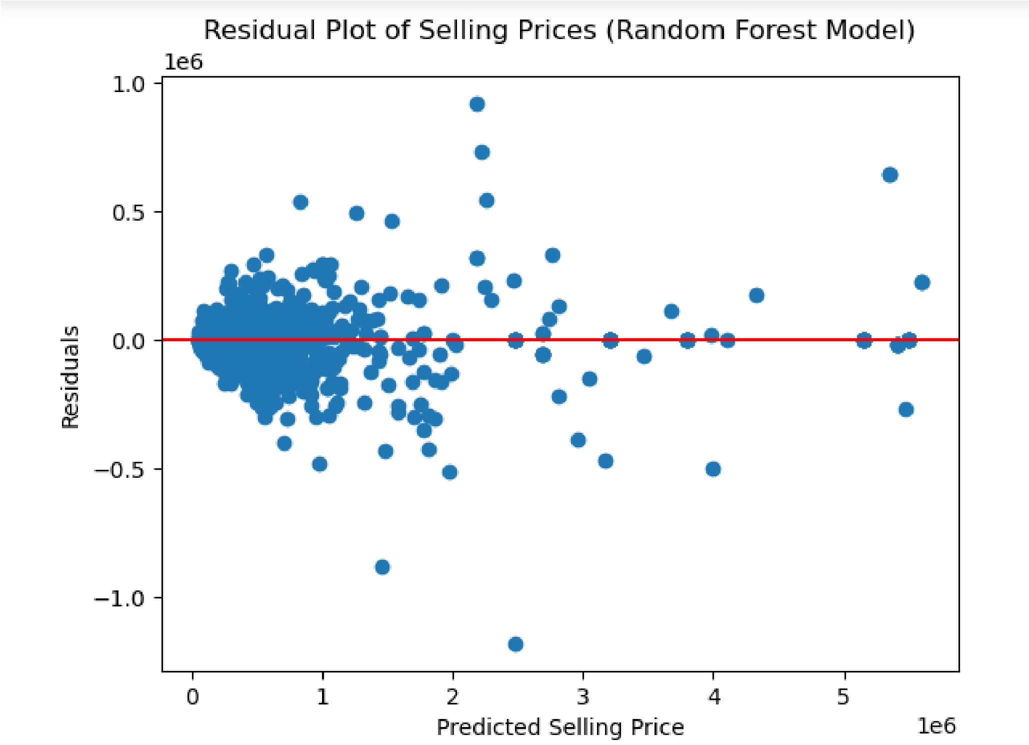
To conclude the project, I showed how a user can pass a predictive dataframe with details for their desired car through the transformative pipeline and utilized the previously trained model to predict a price for the car described in their dataframe. To demonstrate efficiency, the details in my predictive dataframe happened to match one of the cars in the data set, but this is not required.

Visualizations/Analysis

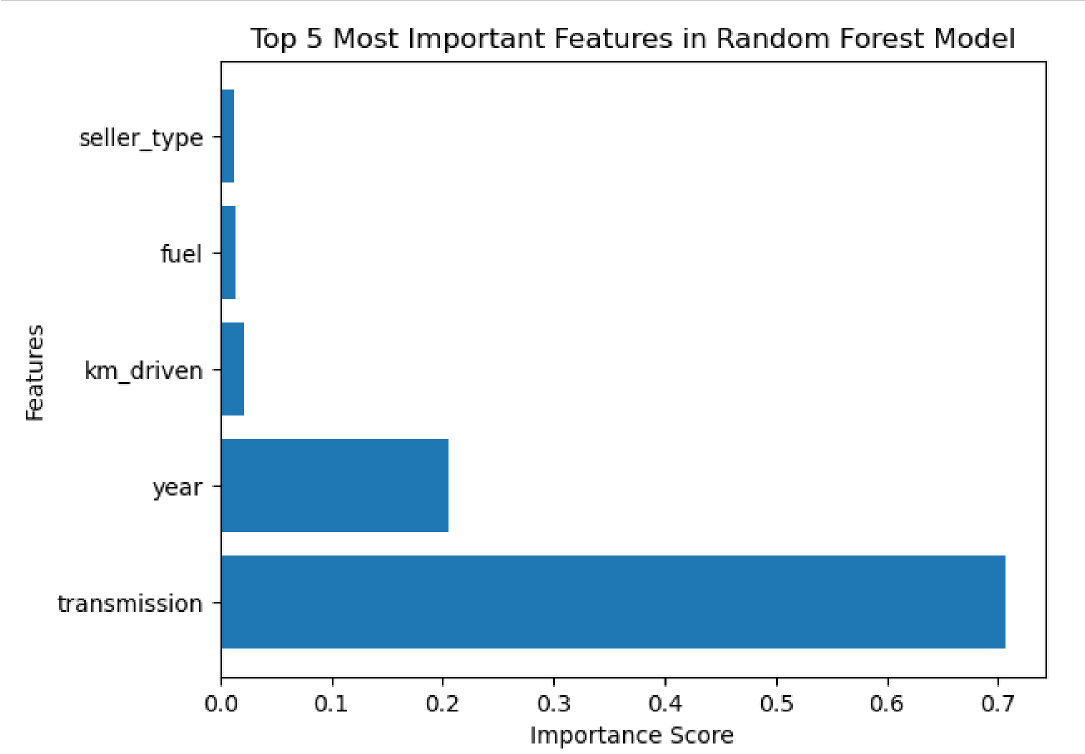
Regression Plot - The below regression plot graphs the predicted sales prices vs the actual sales prices. The plots gather around the red target line indicating a high rate of accuracy with limited variance between actual prices and predicted prices. This also gives a small insight into the spread of the selling price variable. This variable is skewed to the right with more data at the lower end of the price range and outliers on the high end of the price range.



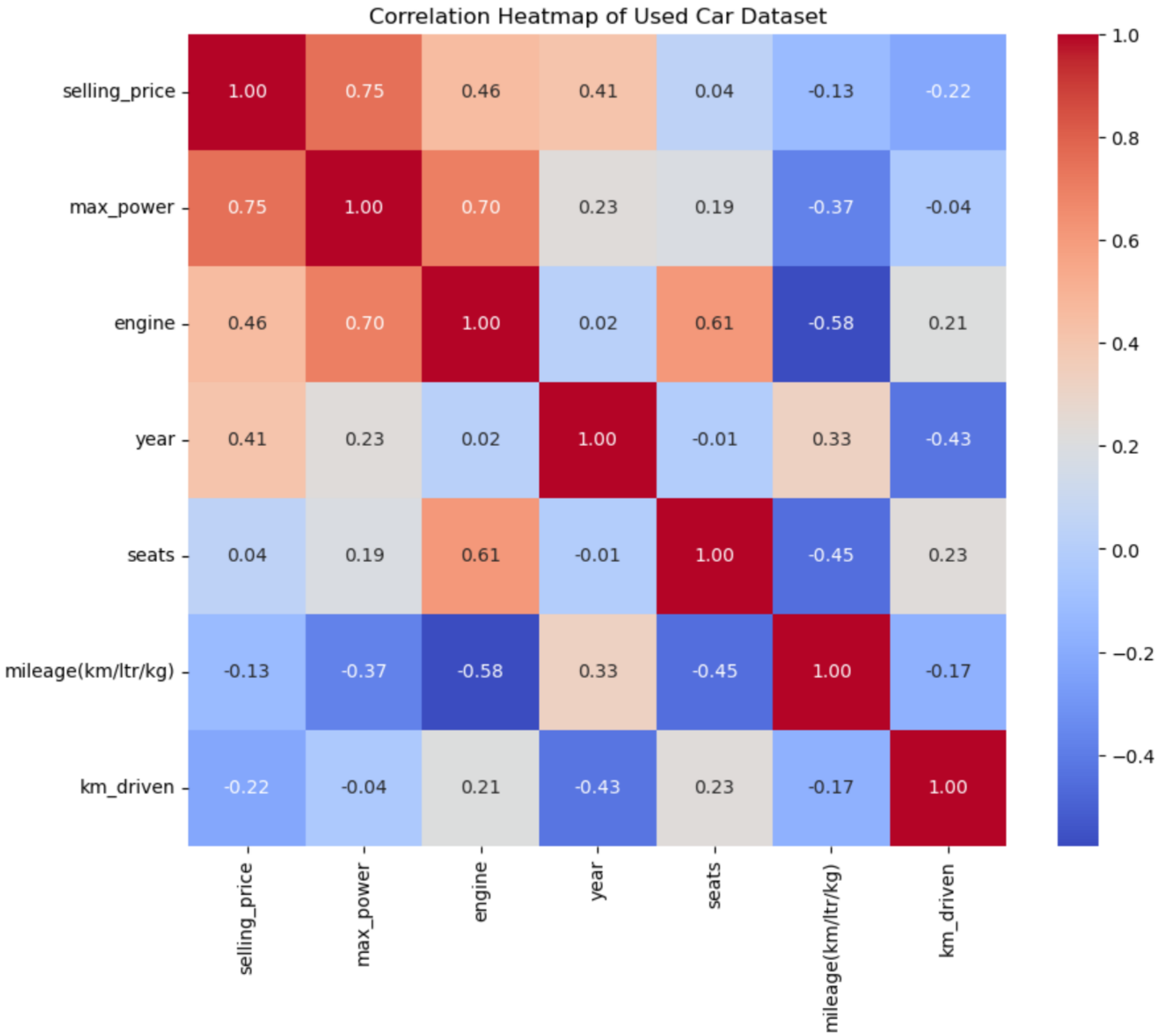
Residual Plot - The below residual plot graphs the residual amounts (the differences between the actual sales prices and the predicted sale prices) vs the values predicted by the model. The plots gather around the y-axis base line of "0" indicating that the difference between the predicted value and actual value is minimal. There appears to be an even distribution of residuals above the actual value and below the actual value.



Features Importance – The bar graph below shows how important the top 5 features are when the random forest model is predicting the selling price. The type of transmission is by far the most important feature followed by year.



Correlation Matrix – The correlation matrix below shows how the numeric variables move in relation to one another. Max_power has a strong positive correlation with selling_price at .75 indicating that max_power and selling_price tend to move together. As max_power increases, selling_price will increase as well and vice versa when the variables decrease. Conversely, km_driven has a -.22 correlation with selling_price. This is a weaker negative correlation, indicating that as km_driven decreases selling_price will increase and vice versa.



Conclusion

In this study, I was successfully able to clean data, use a pipeline to scale numeric columns and make dummies out of categorical columns, and evaluate several regression models based on the R-Squared and RMSE values. I opted to move forward with a Random Forest Regression model. The regression plot and residual plot demonstrate the efficiency of the model with data gathering around the target line (regression plot) and base line of 0 (residual plot). I passed the criteria for my desired car through the pipeline and random forest model and received a predicted selling price. I can utilize the information learned in my correlation matrix and features importance graph to modify my criteria to increase or decrease my selling price. I am happy with the usability and accuracy of my model.

Assumptions

This study assumes that all vehicles in this dataset are in the same area and sold within the same timeframe. In the Kelley Blue Book article "*Average Used Car Price Down 4% Since Last Year",* the author details how factors such as supply and demand and local economics can change the price of a used car. An example provided in the article is from 2020 and 2021 when Covid made the demand for used cars high as production of new cars was waning. At the same time, the supply of used cars was low as drivers were opting to purchase cars at the end of their lease period rather than return the leased cars to the market as a used car (Tucker, 2024.) If some of the cars in the dataset sold during 2020 to 2021, they might not be valid entries in the dataset as today's used car market is different than the market was during Covid. Due to the lack of a "sold date" in this dataset, I am assuming that all the cars were sold in a car market equivalent to the market in which the buyer is currently shopping.

Limitations

The largest limitation of this dataset is that I do not know what unit of currency the selling_price variable is in. Because other variables are measured in kilometers which is not a common unit of measurement in the United States, I do not believe the selling_price is currently listed as U.S. Dollars. If I knew the original unit of currency, I could convert it to dollars. However, that is only a concern with this specific dataset. If a dealership or car broker were to offer this service to their clients, they would be using a dataset of cars where they facilitated the sale and they would be able to identify their unit of currency.

Challenges

My goal for this project is that an individual car buyer can utilize a project such as this to empower them with knowledge during that seems to be a purposely nontransparent car buying process. However, I anticipate it would be a challenge for an individual to obtain a good dataset that has a recent history of used car sales in their locality. Additionally, once a dataset is established, I am not sure if the average individual would be able to replicate the study. I aim to provide step by step instructions in my corresponding presentation, but an individual would still need a base knowledge in data science to successfully replicate this study on their own data set. If a dealership or broker were to replicate this study, they would have their own dataset of cars and can contract the building and training of the model to a specialist.

Future Uses/Additional Applications

Although intended to put the power and knowledge to price a used car directly in the hands of an individual buyer, a car broker or dealership could offer this as a special service to a potential buyer to build trust and add transparency to the process. Alternatively, this study can also be used from a seller's perspective (either an individual seller or a dealership) to ensure that they are pricing their cars fairly and competitively against other cars in the market.

Recommendations

Before deploying this model, it is recommended that an individual create a range for their budget and desired characteristics for the used car they want to purchase. A predetermined range will help the buyer know how far they are willing to alter their characteristics to stay on budget. For example, the budget is 500,000-600,000 and the range for the year that their car was made is between 2010 and 2020. When the predictive data frame is filled in with a year of 2015 and the predicted price is 450,000, the buyer knows they have the budget to update the predictive dataframe to represent a year value of 2020 and include all the benefits of a more modern car.

Once a range has been determined for the budget and the features variables, it is recommended that the buyer creates a predictive dataframe with several different entries representing a variety of cars in which they would be willing to purchase. They can then run the predictive dataframe through the transformation pipeline and random forest model to obtain the predicted prices of the cars described in the predictive dataframe.

Implementation Plan

Once a buyer has input the desired characteristics for several car options in the predicted dataframe and has received the corresponding predicted prices, they will be prepared to utilize the knowledge gained in this study during their car buying journey. It is suggested that they implement this knowledge via the following steps:

1) Review dealerships, online ads, and social media marketplaces for cars that match their desired features.
2) It is unlikely that a car will be found that matches every single feature in their desired predictive dataframe. If a car is found that matches 7 out of 10 desired features, add an entry to the predictive dataframe with the features of the advertised car and determine if the predicted price is close to the listed sale price.
3) I prefer to have negotiations in writing. Therefore, email the dealership/private seller stating that you are interested in the car and based on your calculations you are willing to pay *insert predicted sale price* for the car.
    a. If they agree to this price, set up a time to view and test drive the car.
    b. If they do not agree on this price, use the model as a bargaining tool and explain why your stated price is an acceptable price for the car. This model can be an effective tool for negotiations.
4) View the car in person.
    a. If the car is a good fit, make the purchase knowing that you stayed within your budget and did not leave any money on the table that could have been spent on features that you would have enjoyed. Hopefully, standing firm to the predicted price in the model has helped you avoid high pressure sales tactics.
    b. If you do not like the car and determine you would like to alter some of your desired features (for example you determine you no longer want a Honda and instead would like a Toyota) update your predictive dataframe and start again from step one.

When making the negotiations, bear in mind that while the features in the model do account for 98% of the variance in sales price, the RMSE does note that on average the predicted price is off by 102,669. If a seller is offering a price that is within the range of the RMSE, use your discretion if you would like to negotiate or accept the offer as it is.

Ethical Assessment

I have two primary ethical concerns with this study.

1) ***Dataset* –** If this model is used and deployed by an individual buyer, they will have to ethically obtain a dataset. It is possible that there is a local repository with this data, perhaps via a government office that collects property tax. If there is not a ready-made dataset, the individual will have to create their own dataset possibly by reviewing newspaper ads and social media posts. Whatever dataset they utilize should not have confidential information about the buyers and sellers of the cars as this is privileged information and not pertinent to the features needed to train the model.
2) ***Negotiating* –** No one is required to abide by the predicted price generated by the random forest model. I am confident that the model predicts a fair price within the margin defined by the RMSE, but if a seller insists on valuing their car at a higher price, you cannot force them to lower their value if they are not interested in negotiating. If they are holding firm on their price, you can either walk away or pay a price that would be considered an outlier in the dataset.

CITATIONS

Brar, S. (2024, March 28). *Car price prediction dataset*. Kaggle. https://www.kaggle.com/datasets/sukhmandeepsinghbrar/car-price-prediction-dataset

D'Allegro, J. (2021, October 19). *Just what factors into the value of your used car?*. Investopedia. https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp

Duman, M. (2024). *Ways to save money when creating a used car budget*. Mike Duman Auto Superstore. https://mikeduman.com/blog/ways-save-money-when-creating-used-car-budget

Tucker, S. (2024, February 19). *Average used car price down 4% since last year - Kelley Blue Book*. Kelley Blue Book. https://www.kbb.com/car-news/averaged-used-car-price-down-4-since-last-year/

APPENDIX

Appendix A – Budgeting for a Used Car

Title: "Ways to Save Money When Creating A Used Car Budget"

Publisher: Mike Dunman

Published: 2024

URL: https://mikeduman.com/blog/ways-save-money-when-creating-used-car-budget

Ways to Save Money When Creating A Used Car Budget

How solid is your plan for purchasing a used car? Creating a personal budget to buy a pre-owned vehicle is like budgeting for any major purchase. First, you should determine how much you can afford to pay. How? By reviewing your total household income and expenses and then looking at how much money is remaining. Although this may sound simple to you, many people get caught up in the excitement of shopping for a car and gloss over the expenses involved in the process. Let's discuss ways to save money when creating your optimal used car shopping budget.

Financing Or Cash Payment

A car payment is typically the next most significant bill after your mortgage or rent. Like housing, car expenses come with associated fees such as insurance and maintenance. Don't forget to consider these when creating a car shopping budget. The rule of thumb suggested by financial experts is to spend 10% or less of your net income on the car note and 20% or less on gas, repairs, and all other automobile costs. Aside from the monthly car payments, consider the registration, taxes and possible maintenance costs needed for your pre-owned vehicle at the time of purchase. Based on your calculations, can you afford to pay for your used car in full, or will you need to apply for financing?

If you can afford to pay cash, it probably sounds like a great idea to avoid interest charges. However, if you qualify for a favorable interest rate, you can make a substantial down payment for the automobile and save your cash to purchase other desirable amenities for it. Or you can use it to invest in additional projects. With financing, you can incorporate some of your upfront costs and save future earnings for monthly payments. At the Mike Duman Auto Superstore, you can secure financing easily with the help of a professional team that's dedicated to helping customers with credit approval needs.

Trade In Your Current Car

A quick and simple way to get rid of your current vehicle and save on a used car is to trade it in to a dealership. The amount they give you for it contributes toward the price of another car on their lot. The used car dealer will assess your current car's reliability and value and make you an offer. Upon a mutual agreement, you turn the title over to the dealership. No need for you to market it, show it, sell it, or do the required paperwork with someone else. All of those transactional activities can be handled effectively in one place.

Do you still owe money on your current car? Not a problem. Reputable used car dealerships can easily take care of that also. They can pay off the balance of your loan, obtain the title from the lender, and seal the deal. An additional advantage is positive equity in your current car. That money can be used toward a down payment on your financing or cash purchase. And let's not forget about sales tax savings. Depending on how your resident state designates sales taxes for vehicles, your trade-in value could be deducted from the price of your used car when calculating sales tax.

Other Ways To Increase Savings

There are a handful of other ways to reduce your expenses. By choosing a used vehicle that has a good safety record and a low theft rate, you can decrease your insurance costs. A quick internet search can provide you with the necessary information. Another idea is to eliminate or lower unnecessary current expenses from your budget, such as outside entertainment, restaurant dining, and magazine subscriptions. Pay off a credit card or two to free up money for your wallet. Revise your variable expenses to ensure that you can handle your fixed expenses.

How much should you spend on a used car? Hopefully, you can determine that by reviewing your financial responsibilities and the aforementioned tips for ways to save money.

10 Questions

Question #1: Is there a way to increase the model’s accuracy?

Question #2: What is the benefit of using a pipeline?

Question #3: What if I only know some of the variables in my predictive dataframe?

Question #4: This data has a variable titled km_driven, I would like to review this variable in terms of miles.

Question #5: I want a general idea of how much a used car costs so I understand the volume of money that is involved.

Question #6: There is a huge difference between the min and max values! Can I see the details for the cars with the min and max values so I can understand the wide spread in selling price?

Question #7: Transmission is by far the most important variable in the model based on the Features Importance graph. How does selling price differ between the two transmission types?

Question #8: How does km_driven affect selling_price?

Question #9: I want a car with a powerful engine. Can I see the 10 cars with the most max_power?

Question #10: How does seller_type affect selling_price?