

Slide 1 – Title / Introduction

Hi everyone, today I'll present my analysis on predicting car sale prices using machine learning. The essential question I focused on is: Can we accurately predict a car's price from its features, such as condition, mileage, and year? In addition, I also want to explore two secondary questions: Which features most influence car prices, and how do prediction errors vary across different types of cars?

Slide 2 – The Data

The dataset comes from Kaggle and contains about 560000 observations of car sale data. Each car has numeric and categorical features, such as year, odometer, condition, and state. I decided to hone these categories down into the 9 most convincing and relevant features for the sake of this analysis. Each category is self-explanatory except condition, which seems to rate the condition of the car on a scale of one to one hundred. You can see the first couple rows of the dataset in the image below.

Our target variable is the selling price. Before modeling, I preprocessed the data by encoding categorical variables and splitting it into training and testing sets. This prepares the data for both Linear Regression and Random Forest models.

Slide 3 – Models

Here we compare the performance of Linear Regression and Random Forest models.

I first decided to use Log Regression to establish a baseline analysis of the dataset. You can see that most predictions align well with the actual prices. Overall, the model performed relatively okay, shown by the r-squared of 0.68. One interesting thing to note is that the model predicted some low-cost cars to amounts below zero, showing its weakness at the extreme cost values.

To further analyze performance, I decided to use a Random Forest model to determine which features of the car had the highest impact on its selling price. As you can see by the RMSE and r-squared, the Random Forest also managed to be a better predictor of prices when compared with Log Regression. The same problem arose though, with the Random Forest model also seemingly performing poorly at the cost extremes. However, overall, the Random Forest did extremely well consider how few variables it had to work with.

Slide 4 – Summary

To summarize my learnings, Random Forest outperformed Linear Regression, achieving lower RMSE and higher r-squared. However, both models have relatively high RMSE when considering that most cars sell for under \$50,000.

Looking at feature influence, odometer, make, and body type were the most powerful predictors of price, while transmission, interior material, and color had minimal impact. The top three contenders logically make sense, as these do impact the production cost and therefore final

selling price, but I was shocked to see that interior material had such a low impact despite the large differences between common interior materials like foam vs leather.

Focusing on Random Forest residuals as the better overall model, the mean residual shows the model slightly underestimates prices, but it is very close to zero overall. The standard deviation is high, reflecting that extreme prices contribute to larger errors, while the MAE is lower, indicating typical predictions are fairly accurate and less affected by outliers.

In conclusion, yes, we can predict car sale prices. Both models we used were relatively successful despite the small number of variables. Odometer, make, and body type, in accordance with common sense, do have the greatest impact on price. And finally, the Random Forest model clearly won out in effectiveness in providing robust analysis as well as in predictive performance in general.