TECHNICAL REPORT

Predicting Car Stopping Distance

Tanja Adzic

Introduction

The goal of this project is to predict a car's stopping distance by constructing a predictive model that will utilize the car's speed and the corresponding stopping distance. Interestingly enough, the data used for this project dates from the 1920's, and I believe it still provides a solid understanding of the speed-stopping distance relationship that remains valid in today's automotive industry.

In order to employ a successful performance of machine learning algorithms, there are a couple of phases needed to be taken. First, I will use basic dataset exploration that will help me understand the data and the relationship between these two attributes. Moving forward to the next phase, the visualization part, which will help me greatly in visually inspecting the attributes and the data points. The phase of data modeling will deal with algorithms, so called regression algorithms, that are tasked to find the best fit to data, which will then be evaluated using suitable metrics. This evaluation phase will show me how well the algorithms performed, and predicted the stopping distance.

Successful implementation of a good performing model could have an impact on road safety. By equipping drivers with a tool to anticipate stopping distances based on their car's speed, there is potential to reduce accidents caused by insufficient braking distances. This not only enhances the safety of drivers and passengers but it also contributes to the overall well-being of pedestrians.

Data

The source of the dataset is Ezekiel, M. (1930) Methods of Correlation Analysis. Wiley. They note the dataset is from the decade of the last century - the 1920's.

The dataset has two numerical attributes:

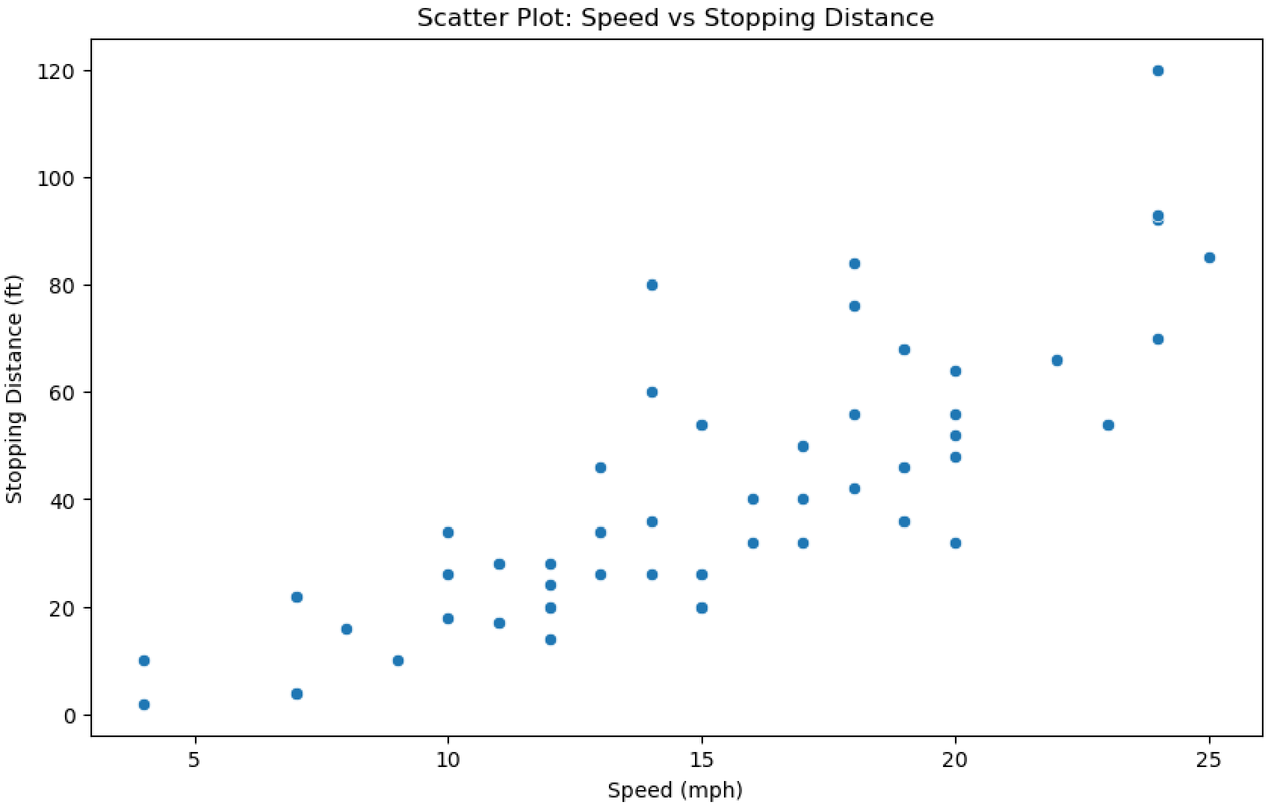
- *speed* - speed of the car in mph
- *dist* - stopping distance in feet

Dataset Exploration and Visualization

Initial inspection of the dataset showed no anomalies in regards to data types, data structure, there were no missing values or duplicates. I was able to inspect the entire dataset visually as it consisted of 50 points. There were also no apparent outliers in the data. The dataset showed a consistent and coherent structure. Data types were aligned with their intended representation and given description.

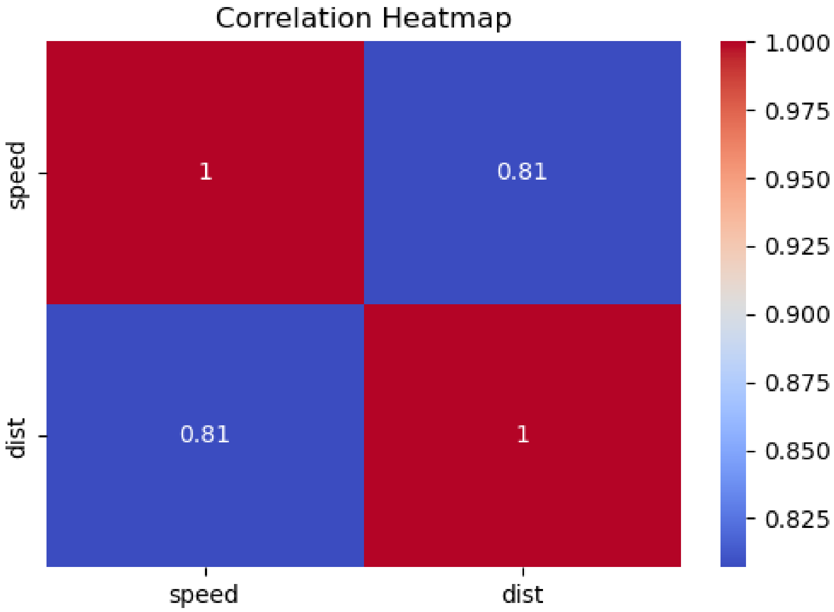
I performed data visualization in order to gain insights into the relationship between the speed and the distance attributes.

What is the relationship between vehicle speed and stopping distance?



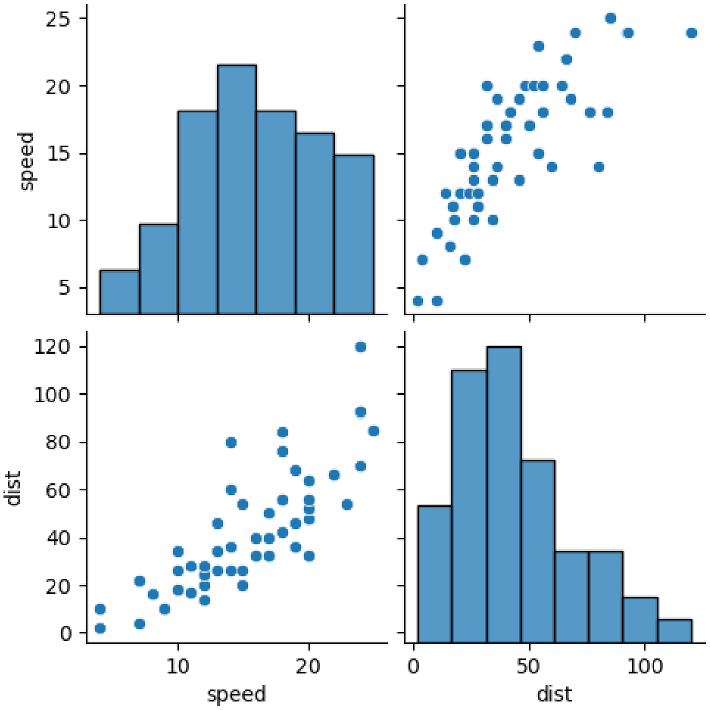
I instantly noticed a linear upward trend between speed and distance. The positive trend indicates that as the speed increases, the stopping distance tends to decrease. According to the data provided, and plotted, higher speeds require more distance to stop safely.

How are speed and stopping distance correlated?



As previously suggested, this plot shows a positive correlation between car speed and its stopping distance. This means that as one attribute increases, the other attribute tends to increase as well.

What is the distribution of data points?



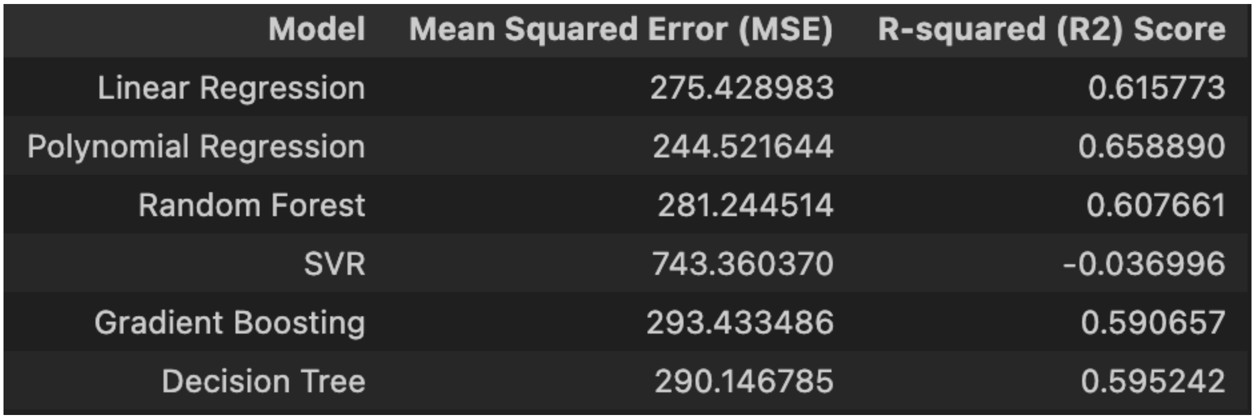
The histograms show how data points are distributed. It appears that the majority of data points have speeds around 10 to 20 mph, and that they are roughly symmetric. On the other hand, most of the stopping distances are clustered around the lower values, with a few outliers at higher distances, which indicates a negatively skewed distribution, and the mean is pulled to the left.

Methodology

In this data analysis, I chose various modeling techniques in order to understand and predict the relationship between the variables. I used linear regression to create a baseline model for predicting stopping distance based on speed. Random Forest, a powerful ensemble algorithm, was utilized to capture complex interactions and improve prediction accuracy. Additionally, I also used polynomial regression to account for potential nonlinear relationships, even when the data showed a steady positive trend. I also used Support Vector Regression, Gradient Boosting, another ensemble method, and a Decision Tree model that I implemented to see any hierarchical patterns within the data. I used these methodologies to try a range of approaches, from linear to nonlinear and ensemble techniques, providing a comprehensive analysis of the dataset and enhancing prediction capabilities.

Data Modeling and Evaluation

In this data modeling and evaluation process, as I stated in the methodology section, I used a variety of machine learning algorithms to predict the stopping distance based on speed. The models I employed included Linear Regression, Polynomial Regression, Random Forest, Support Vector Regression (SVR), Gradient Boosting, and Decision Tree. I trained and tuned each model using grid search over relevant hyperparameters. As a performance evaluation, I used metrics such as Mean Squared Error (MSE) and R-squared. The results revealed valuable insights about the models' predictive capabilities:

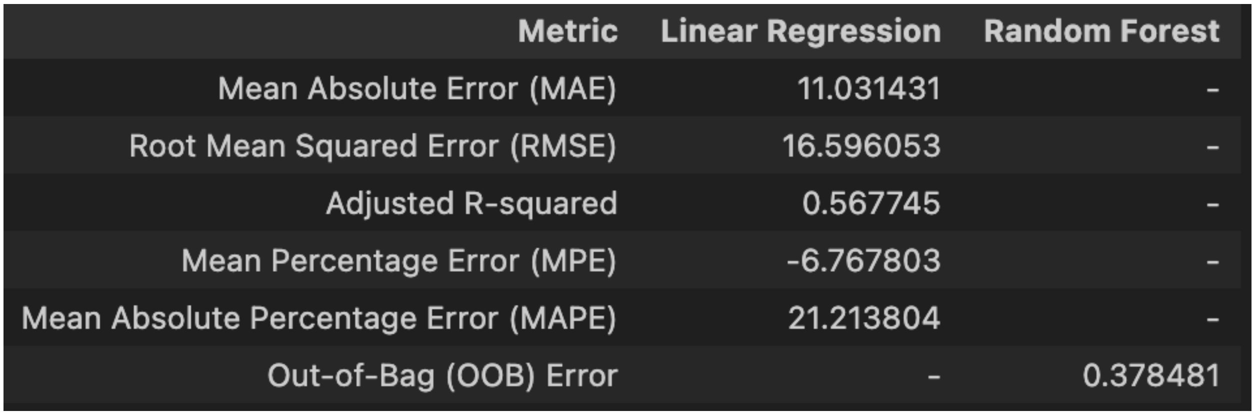


Model	Mean Squared Error (MSE)	R-squared (R2) Score
Linear Regression	275.428983	0.615773
Polynomial Regression	244.521644	0.658890
Random Forest	281.244514	0.607661
SVR	743.360370	-0.036996
Gradient Boosting	293.433486	0.590657
Decision Tree	290.146785	0.595242

The Polynomial Regression exhibited the lowest Mean Squared Error (MSE) of 244.52 and a valid R-squared score of 0.66, suggesting that its quadratic nature captured some of the inherent nonlinearity in the data. Linear Regression, Random Forest, Gradient Boosting, and Decision Tree followed closely, with MSE values ranging from 275.43 to 290.15 and R2 scores ranging from 0.59 to 0.62. These models also seem to have effectively captured relationships within the dataset, however when it comes to nonlinearity the results were inconclusive.

On the other hand, the Support Vector Regression (SVR) model displayed significantly higher MSE (743.36) and a negative R2 score (-0.04). This suggests that the SVR model did not fit the data accurately. The performance of the SVR model could be improved with further parameter tuning or feature engineering.

Additionally, I added a couple of specific metrics for Linear Regression and Random Forest to gain further insights:



Metric	Linear Regression	Random Forest
Mean Absolute Error (MAE)	11.031431	-
Root Mean Squared Error (RMSE)	16.596053	-
Adjusted R-squared	0.567745	-
Mean Percentage Error (MPE)	-6.767803	-
Mean Absolute Percentage Error (MAPE)	21.213804	-
Out-of-Bag (OOB) Error	-	0.378481

The Mean Absolute Error (MAE) for Linear Regression is approximately 11.03, indicating an average prediction error in the stopping distance of 11.03 feet. The Mean Absolute Percentage Error (MAPE) is 21.21%, implying that, on average, the model's predictions deviate by around 21.21% from the actual values. The Adjusted R-squared value of 0.57 for Linear Regression suggests that the model explains about 57% of the variability in the data, which I deemed as a moderate fit. The Out-of-Bag (OOB) Error for Random Forest is 0.378 that shows the model's performance on unseen data.

The Polynomial Regression appears to be the most accurate among the tested models, with Linear Regression, Random Forest, Gradient Boosting, and Decision Tree performing reasonably well. The SVR model, however, struggled to capture the complexities of the data.

Model Limitations and Future Recommendations

While the models I employed in this project show promising predictive performance for the stopping distance based on speed, there are certain limitations that I need to address. Firstly, I believe the dataset is limited in size and scope, and that can potentially lead to overfitting in more complex models. Additionally, the assumption of linearity in the Linear Regression and Polynomial Regression models may not hold true for all scenarios, as real-world relationships can be more intricate. Finally, the Support Vector Regression (SVR) model demonstrated poor performance, indicating its sensitivity to hyperparameters and the need for further fine-tuning.

I recommend expanding the dataset with a broader range of attributes like driving conditions, road types, and weather conditions, as well as vehicle attributes such as weight to improve the models' generalization capabilities. Incorporating additional features like vehicle weight, road conditions, and driver behavior could also contribute to the models' accuracy. I believe experimenting with more advanced techniques, such as neural networks or ensemble methods, could also yield more accurate predictions by capturing complex interactions in the data.

Conclusion

Throughout this project my aim was to explore a range of regression and ensemble models to uncover the relationship between car speeds and their stopping distances, and provide insights for better decision-making. I recognized that Linear Regression and Polynomial Regression models showed promising results, suggesting potential nonlinear trends in the data. Additionally, Random Forest and Gradient Boosting models showcased their ability to capture complex relationships. However, I acknowledged that the Support Vector Regression (SVR) struggled to generalize, possibly due to its sensitivity to parameters and data distribution.

In evaluating the models, I utilized metrics like Mean Squared Error (MSE), R-squared scores, and additional statistics to investigate their performance comprehensively. Considering the limitations of a relatively small dataset, assumptions of linearity, and the scope of features, I propose future enhancements such as dataset expansion, inclusion of additional features, and exploration of advanced techniques to further improve predictive accuracy. This project underscores the importance of data-driven approaches in road safety analysis and aims to contribute to enhanced safety awareness and informed decision-making.