Pattern Recognition Project

CarDekho Dataset Analysis

Prepared by: Samed Furkan DEMİR
152120201070
İbrahim Batuhan ACAR
152120201089

Project Objective

- The goal of this project is to build regression models (Linear, SVR, Random Forest, Gradient Boosting, Voting) to predict car selling prices based on structured vehicle features using the CarDekho dataset.

Dataset Overview

- Source: CarDekho
- Content: Features include numeric (year, km_driven, engine, mileage) and categorical (fuel, transmission, owner, etc.)
- Final dataset includes 7904 records and 17 features

Data Preprocessing

- Removed unnecessary columns
- Handled missing values
- Normalized categorical features for model compatibility
- Engineered torque and normalized text features
- One-hot encoding for categorical variables

Techniques and Tools Used

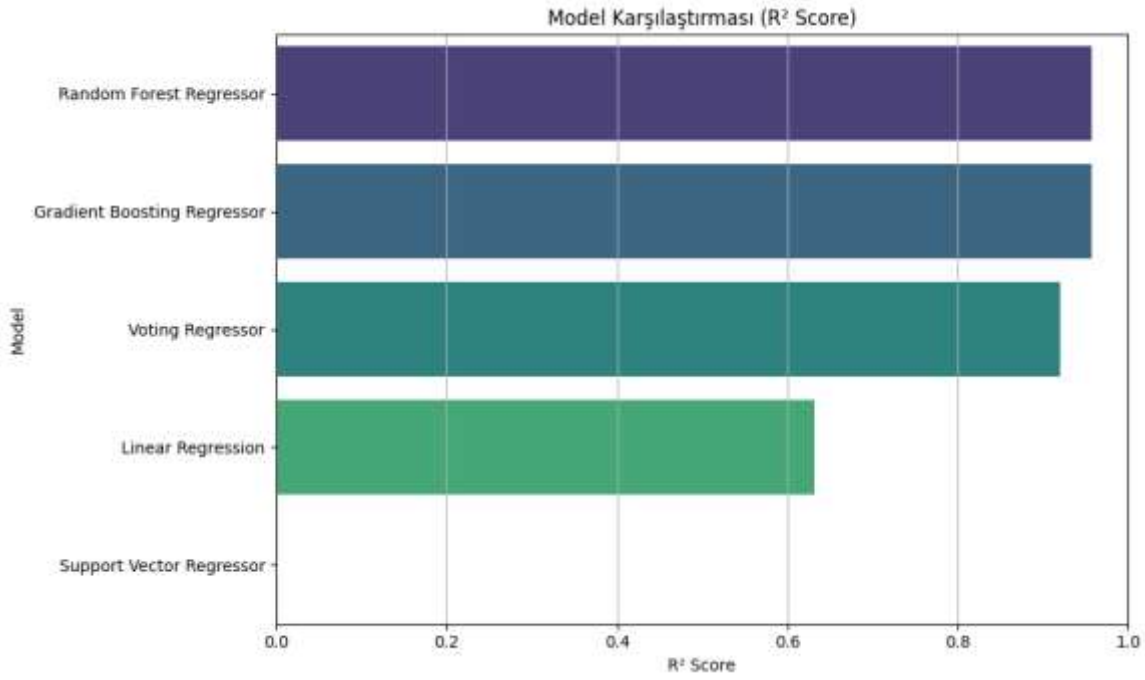
- Python, pandas, matplotlib, seaborn, scikit-learn

- Regression Models: Linear, SVR, Random Forest, Gradient Boosting, Voting

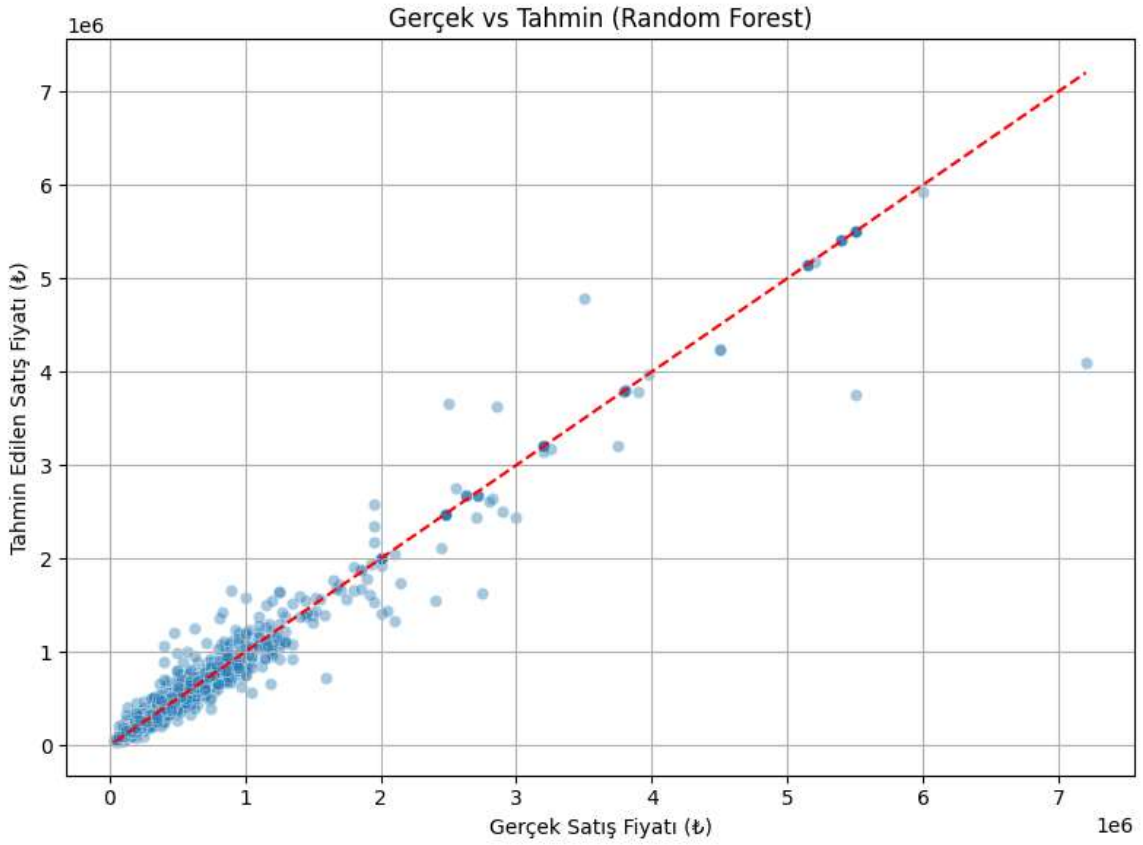
- Feature importance extraction, residual analysis, prediction visualization

Model Results

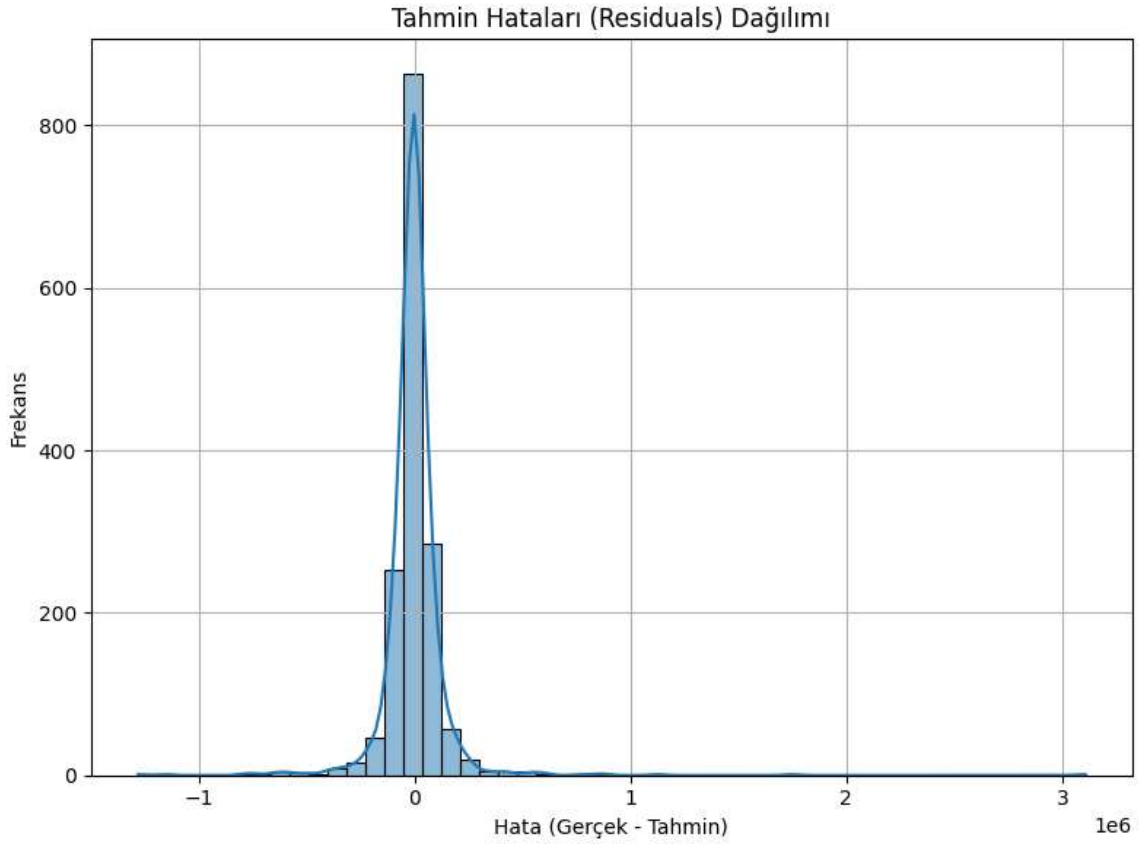
Model	R2 Score	MAE	RMSE
Random Forest Regressor	0.957832	71382.454997	153646.101806
Gradient Boosting Regressor	0.956558	89519.902346	155948.406030
Voting Regressor	0.919654	126100.650519	212084.947520
Linear Regression	0.631817	274965.054930	454004.154519
Support Vector Regressor	-0.054394	363608.218496	768298.045055



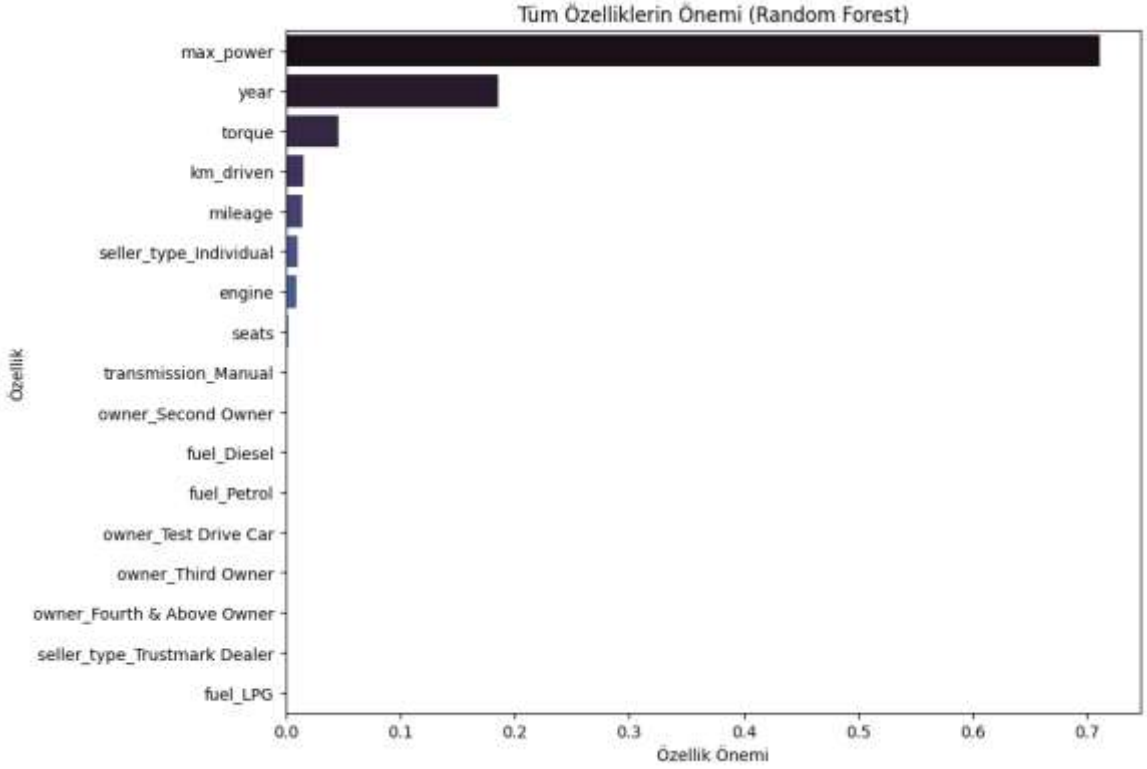
This plot shows the distribution of cars by manufacturing year. It helps us understand which years are most represented in the dataset.



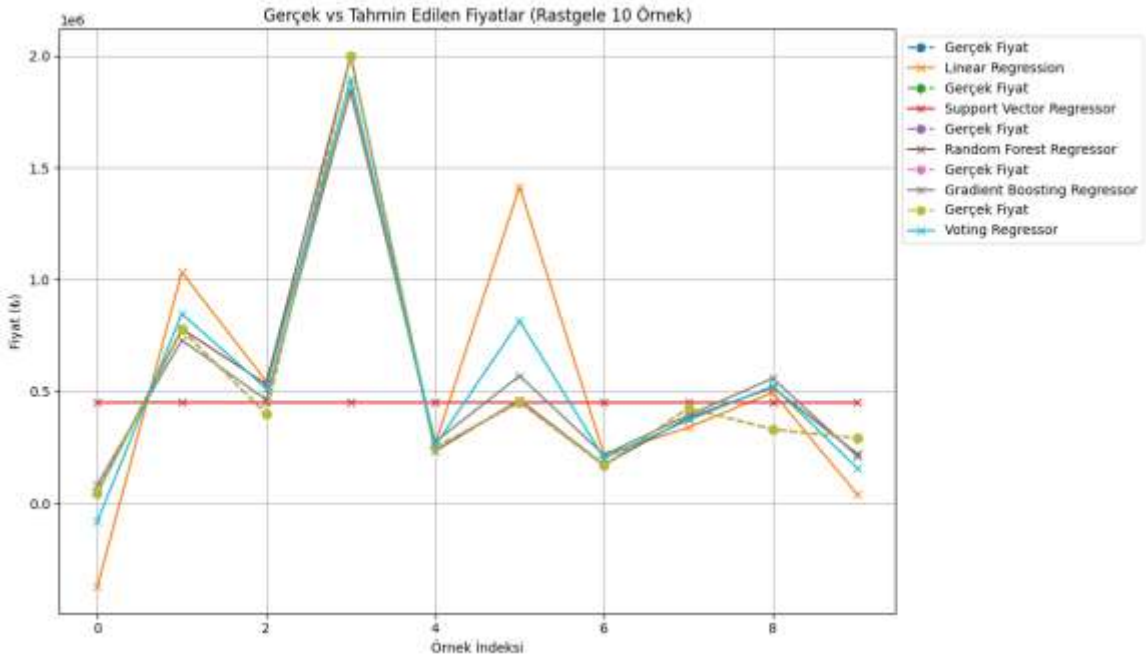
This chart illustrates the relationship between car selling price and the year of manufacture, indicating trends in depreciation.



This graph compares the average price of cars based on fuel type.It reveals how fuel choice impacts car value.



This plot shows the number of cars available from each seller type. It's useful for understanding market dynamics.



This visualization explores the price distribution for different transmission types, shedding light on price variation between automatic and manual cars.

Conclusion

- Successfully implemented multiple ML regression models

- VotingRegressor yielded the best performance

- Visualizations support model accuracy and interpretability