Predictive Analysis: Damage Prediction Using Machine Learning

Methodology

1. **Data Merging and Feature Engineering:** This involved combining datasets to create a dataframe with features like age, injuries, vehicle details, and damage. New features were also engineered to enhance predictive performance.
2. **Target Variable and Features:** avg\_damage\_per\_crash was used as the target variable to be predicted by models. The selected features included avg\_age, avg\_speed\_limit, total\_injuries, severe\_injuries, avg\_vehicles, and avg\_people.
3. **Data Preparation:** The dataset was divided into training and testing sets. Feature scaling was performed specifically for the KNN algorithm, which is sensitive to the scale of the features.
4. **Model Selection:** Three machine learning models were used: Decision Tree, KNN, and Random Forest.
5. **Model Training and Evaluation:** Each model was trained using the training data and evaluated using the testing set. Mean Squared Error (MSE) was used to measure the performance of each model.
6. **Feature Importance Analysis:** For the Random Forest model, feature importances were calculated to understand which features were most relevant in predicting the target variable.

Results

* **Model Performance:**
  + Decision Tree MSE: 1550635983.24
  + KNN MSE: 1212807231.04
  + Random Forest MSE: 713679904.78
* **Feature Importance:** The bar chart showing feature importance revealed that total\_injuries and severe\_injuries were the most important features in predicting avg\_damage\_per\_crash. avg\_people, avg\_vehicles, and avg\_speed\_limit also contributed to the model's predictions.

A graph with blue bars

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

A graph with blue and orange bars

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

A graph with a line

AI-generated content may be incorrect.

Analysis and Discussion

* Random Forest outperformed the other models with the lowest MSE, implying greater accuracy in damage prediction.
* The importance of total\_injuries and severe\_injuries underscores the intuitive relationship between severity of crashes and the associated damage costs.
* The correlation matrix could offer further insights into feature relationships, guiding model improvement.

Time Series Analysis: Forecasting Total Damage Using ARIMA, SARIMAX, and Exponential Smoothing

Methodology

1. **Data Aggregation:** Crash data was aggregated by 'Date' and 'Damage' to create a time series representing the total damage over time.
2. **Model Selection:** Three models were used: ARIMA, **SARIMAX**, and Holt-Winters Exponential Smoothing. The **SARIMAX model was chosen for its ability to incorporate exogenous variables (external factors)** that could influence the target variable (total damage).
3. **Model Training and Forecasting:** The models were trained on historical data and used to forecast damage values into the future. **For SARIMAX, relevant exogenous features were selected and included in the model.**
4. **Visualization:** The historical damage, forecasts, and confidence intervals of the forecasts were visualized on a plot for analysis.

Results

* ARIMA, **SARIMAX**, and Exponential Smoothing models provided forecasts for the next 12 months of the time series.
* **The SARIMAX model leveraged exogenous variables to potentially enhance prediction accuracy.**
* The plots visualized the historical data alongside the forecasted values, as well as the confidence intervals around the predictions.

A graph showing a line graph

AI-generated content may be incorrect.

A graph showing a line graph

AI-generated content may be incorrect.

Overall Conclusions

Based on the analyses, Random Forest was identified as a promising model for predicting damage costs associated with crashes. The models, including **SARIMAX**, were able to forecast city-wide total damage over time with acceptable accuracy. **SARIMAX, with its ability to leverage exogenous factors, offers the potential for more nuanced and potentially more accurate forecasting.**