Story of Our Preprocessing Journey

Initial Idea

Our initial dataset presented us with a diverse range of incident severity types. We identified 30 distinct severity types categorized under four main categories:

- **ACCELERATING**: HA1, HA2, HA3
- **BRAKING**: HB1, HB2, HB3
- **HARSH CORNERING**: HC1 to HC21 (excluding HC9 and HC12)
- **SPEEDING**: SP1 to SP5

Our plan was to build a model to predict the `incident_severity` feature. Given the multi-class nature of the classification task, we initially considered using conventional models like Random Forest or Linear Regression to classify the incidents into categories such as Braking, Harsh Cornering, or Speeding. However, we were aware of the challenges these models face when dealing with a large number of classes, as they are not inherently designed to handle multi-class classification efficiently.

To address this challenge, we proposed creating a new numerical feature that would mirror the `incident_severity` feature with numerical values. For instance, replacing the `HC11` severity type with the number 11 would indicate the intensity of the incident in a numerical form. This transformation would facilitate the use of conventional models by reducing the complexity of the classification task.

Transition to a Better Dataset

As we progressed, we discovered a better dataset that included additional valuable features, which promised to enhance our analysis significantly. The columns in this new dataset, as shown in the screenshots, covered various aspects relevant to our analysis:

- **Town**
- **First Mode of Transport**
- **Second Mode of Transport**
- **Area Type**
- **Light Condition**
- **Road Location**
- **Road Condition**
- **Road Surface**
- **Road Situation**
- **Speed Limit**
- **Street**
- **Weather**
- **Accidents**

This dataset provided a comprehensive view of each incident, capturing essential variables that could influence the outcome of an accident. We began by plotting the data to identify any missing values. Our plots revealed that there were no missing values in this dataset. However, we noticed that many columns contained the value "UNKNOWN." This insight was crucial as it indicated potential gaps in the data quality or issues with data collection.

Addressing "UNKNOWN" Values

To tackle the "UNKNOWN" values, we considered their distribution across different columns. If "UNKNOWN" values dominated a column, it suggested that the data for that attribute might be unreliable. However, before deciding to drop any columns, we explored combining the dataset with additional geographical data, specifically latitude and longitude information. This approach aimed to enrich the dataset by providing precise location information for each incident.

Data Integration and Feature Selection

By integrating the new dataset with latitude and longitude data, we could pinpoint the exact locations of the accidents. This integration was achieved using an inner join, which ensured that only records with matching location data were included in the final dataset. This method enhanced the dataset's richness, enabling a more detailed analysis of accident patterns based on geographical locations.

Given that the new dataset encompassed all necessary features, including weather incidents and other relevant information, we decided to discard the initial dataset. Our focus shifted to predicting whether an incident would result in fatal injury or material damage based on the features available in the new dataset.

Data Cleaning and Preparation

With the combined dataset, we embarked on a thorough data cleaning process. We ensured there were no missing values or inconsistencies that could skew our analysis. The presence of "UNKNOWN" values was addressed by either imputing reasonable estimates based on other available data or by removing records where such imputation was not feasible.

Handling Outliers

Outliers in the dataset were identified and handled carefully. Outliers could significantly impact the performance of predictive models, especially in a dataset dealing with accident severity. We used statistical methods to detect and manage outliers, ensuring that they did not distort the model's predictions.

Normalization and Scaling

To ensure the data was on a level playing field, we normalized and scaled the features. This step was crucial because features measured on different scales could bias our models. Techniques like min-max scaling and standardization were applied to adjust the values so that each feature contributed equally to the analysis. This was particularly important for algorithms sensitive to the scale of data, such as support vector machines and k-nearest neighbors.

Encoding Categorical Variables

The dataset contained several categorical variables that needed to be converted into a numerical format for machine learning algorithms to process. We used various encoding techniques, including one-hot encoding for nominal variables and ordinal encoding for those

with an inherent order. This conversion allowed the categorical data to be seamlessly integrated into our models.

Feature Engineering

Feature engineering was a critical step in our preprocessing journey. We scrutinized the existing features and pondered how new features could be derived to better capture the underlying patterns in the data. This involved creating interaction terms, polynomial features, and aggregating data at different granular levels. Feature engineering required domain knowledge and intuition, transforming raw data into meaningful predictors.

Splitting the Data

Before diving into model building, we made an essential split in our dataset: separating it into training and testing sets. This split was vital to ensure that our models would be evaluated on unseen data, providing a realistic estimate of their performance. We used stratified sampling techniques where necessary to maintain the distribution of target variables across the splits, ensuring that the training and testing sets were representative of the overall dataset.

Conclusion

Through meticulous preprocessing, including data integration, cleaning, feature selection, and engineering, we prepared a robust dataset for our predictive modeling task. Our journey highlighted the importance of data quality, the value of integrating supplementary data, and the need for thoughtful feature engineering to improve model performance. With this well-prepared dataset, we are now poised to build models that can accurately predict the severity of incidents, providing valuable insights for enhancing safety measures and reducing accident risks.