PREDICTING RATES OF CAR COLLISIONS

BY AUGUST STAPF

CONTENTS

- The process of data analysis
- Techniques for analyzing quantitative data
- Analytical activities of data users
- Barriers to effective analysis
- Practitioner notes

THE PROCESS OF DATA ANALYSIS

- The data was originally over 190,000 rows and had 37 attributes
- Pre-processing included picking the significant variables, standardizing the data, and converting categorical variables into binary options.
- After pre-processing, the data shape was converted to 155,000 rows and 32 attributes, which included many binary variables.

TECHNIQUES FOR ANALYZING QUANTITATIVE DATA

- I utilized 3 machine learning algorithms in order to train and test models and determine the most accurate model
  - K Nearest Neighbors
  - Decision Trees
  - Logistic Regression

ANALYTICAL OUTCOMES OF THE APPROACHES

- Accuracy was measured by the jaccard prediction accuracy score
  - KNN: 0.719815
  - Decision Tree: 0.742468
  - Logistic Regression: 0.727854
- Decision Tree had the highest accuracy at almost 75%, which represents the significance of the attributes

BARRIERS TO EFFECTIVE ANALYSIS

- The data was sparse in how the severity code was measured
- Severity code usually has 5 possibilities, but the data only included two of the possibilities

PRACTITIONER NOTES

- Thankyou to IBM for this great opportunity!