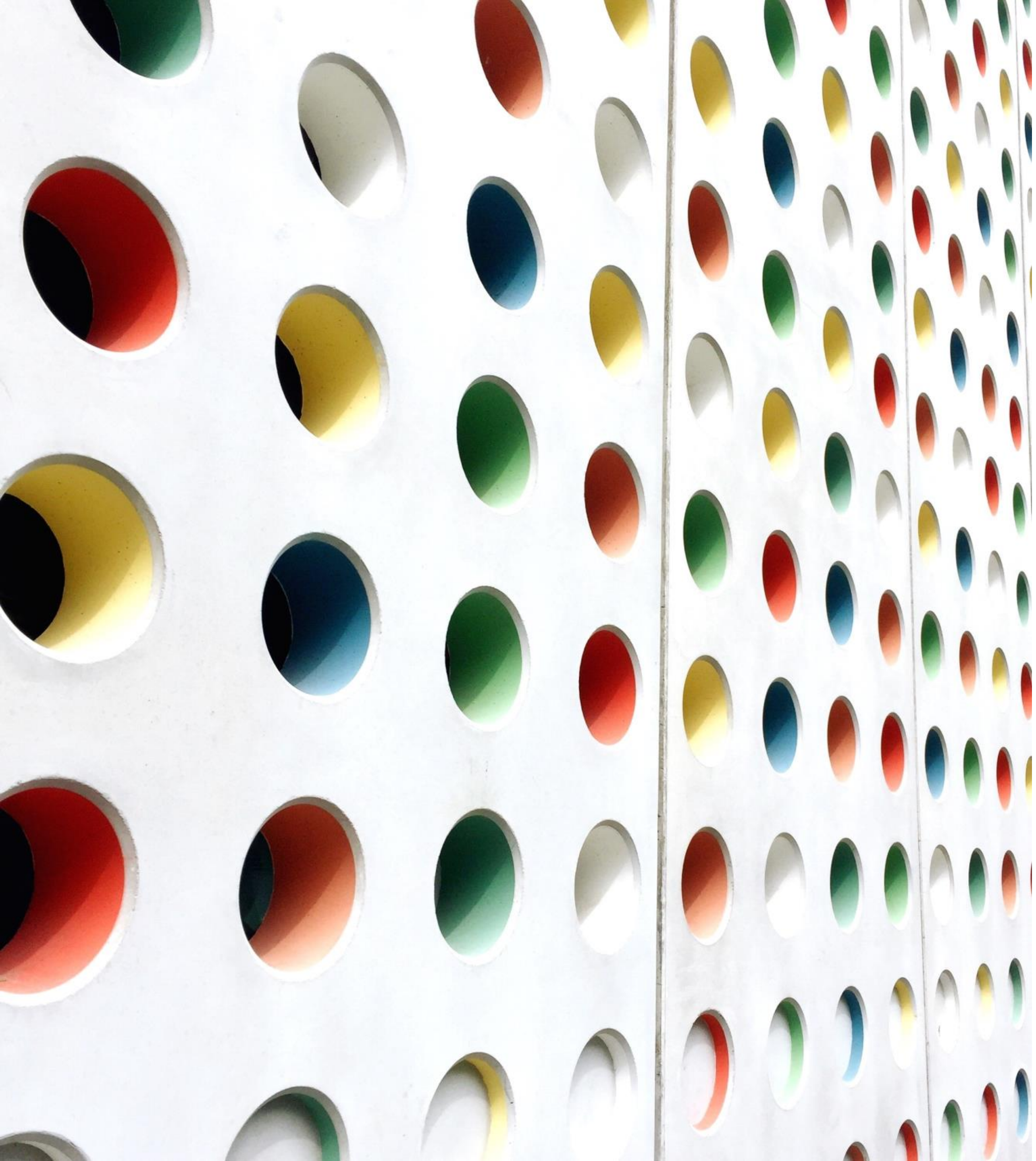
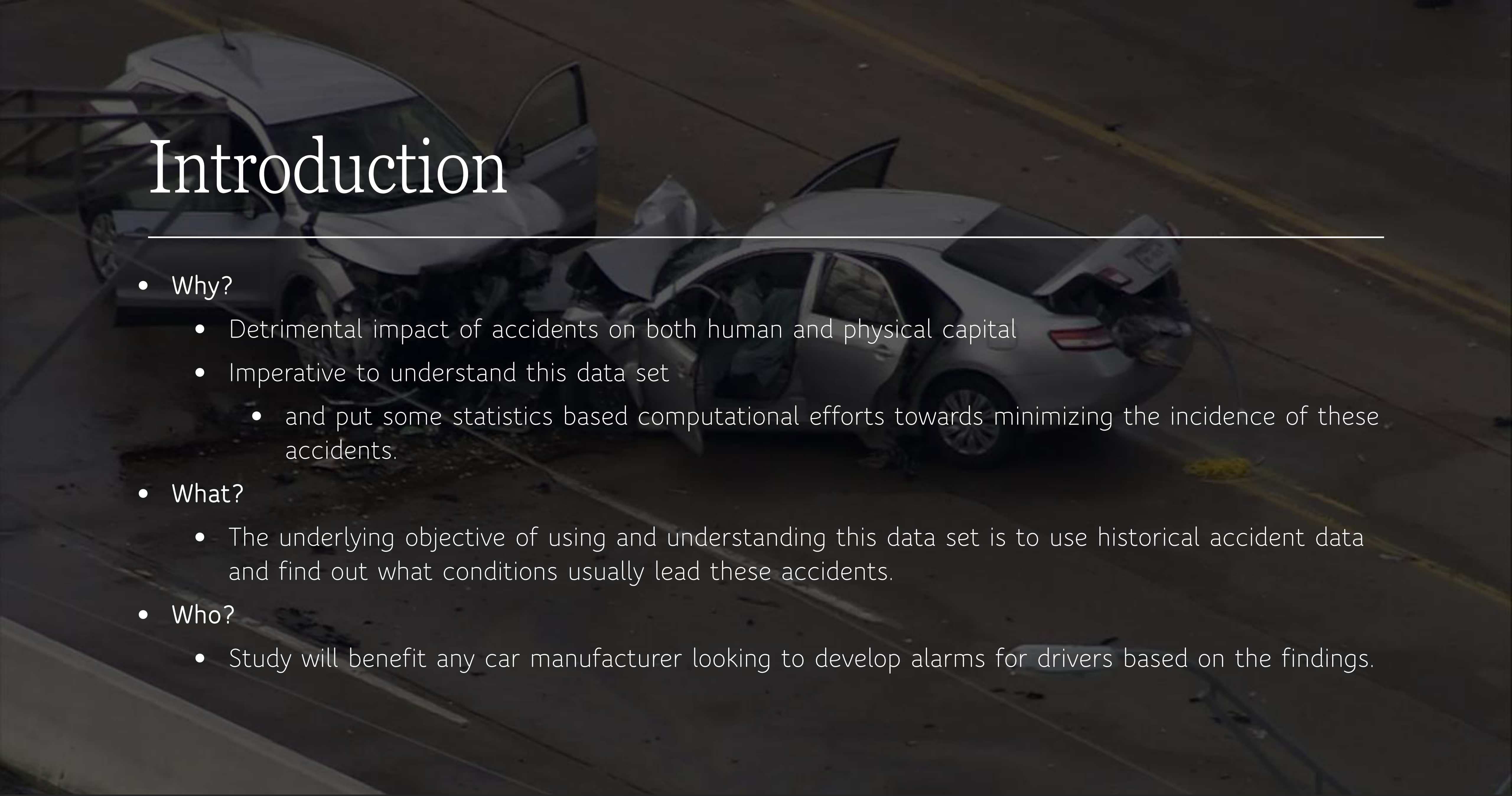


IBM Data Science Capstone

ACCIDENT SEVERITY PREDICTION

Introduction



- Why?
  - Detrimental impact of accidents on both human and physical capital
  - Imperative to understand this data set
    - and put some statistics based computational efforts towards minimizing the incidence of these accidents.
- What?
  - The underlying objective of using and understanding this data set is to use historical accident data and find out what conditions usually lead these accidents.
- Who?
  - Study will benefit any car manufacturer looking to develop alarms for drivers based on the findings.

Data

- How?
  - The data set consists of all collisions provided by the Seattle Police Department and recorded by Traffic Records.
  - Includes all types of collisions.
  - Time frame of the data collected is from 2004 to present.
  - Data broadly covers the time, place and conditions under which these collisions took place. The severity code is described as follows:
    - 3 – Fatality
    - 2b – Serious injury
    - 2 – Injury
    - 1 – Property damage
    - 0 – Unknown

Methodology

Step 1: Data Loading and Cleaning

Pandas Data frame, Remove unnecessary data and drop rows with missing data

Step 2: Data Pre-Processing for Model Building

Encode Categorical Variables, Balance dataset, divide into feature set and target variable set, Standardize data and divide into training, test data

Step 3: Model Building and Evaluation

KNN, Decision Tree Classifier, Support Vector Machine and Logistic Regression

Results

- Model Evaluation
  - Jaccard Similarity Score
    - KNN = 0.556
    - Decision Tree = 0.559
    - SVM = 0.559
    - Logistic Regression
  - F-1 Score
    - KNN = 0.520
    - Decision Tree = 0.531
    - SVM = 0.530
    - Logistic Regression = 0.530
  - Log Loss
    - Logistic Regression: 0.665

Conclusion

Based on the model evaluation, it was determined that the Logistic Regression is the most suitable model to predict the accident severity with a Log Loss score of 0.66. This was also expected because the dataset only consisted of two severity code outcomes, and based on that binary output result condition, a Logistic Regression is the best suited model. The model still needs further finetuning to get the best possible accuracy

It is evident that the road, weather and light conditions are extremely important while driving. This study basically shows a way to predict accident severity ideally using a logistic regression model to prevent or reduce the severity of accidents in the future based on historical data