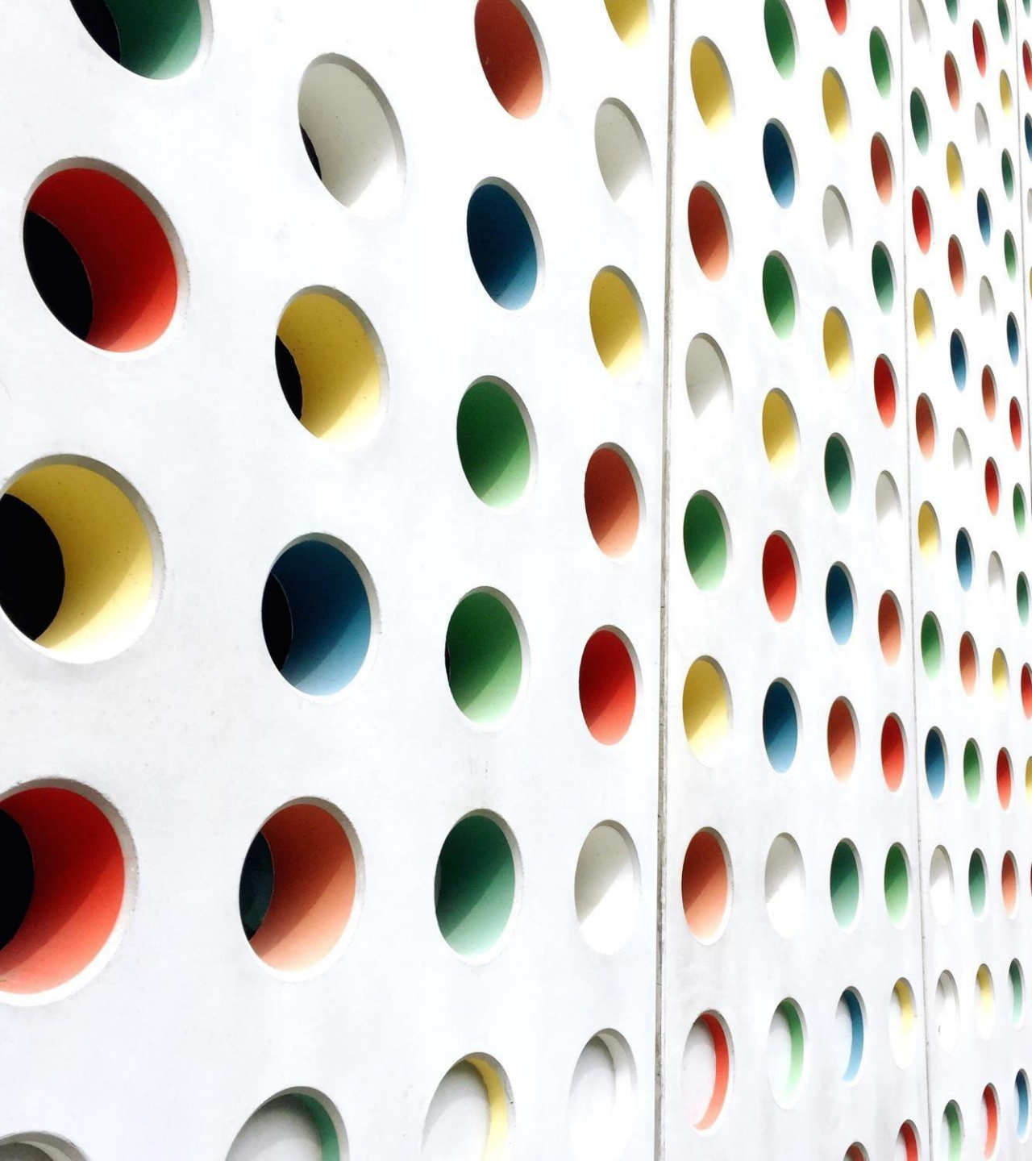


Predicting Car Accident Severity

COURSERA PROJECT CAPSTONE

By Mohd Zunaid

Introduction

- Road traffic injuries are predicted to become the seventh leading cause of death by 2030 and according to the WHO, 1.35mil people die each year
- To attempt to reduce the frequency of car accidents, a model can be developed to predict the severity of an accident given the current weather, road and visibility conditions, whether the driver was distracted and/or under the influence
- This predictive model can be used by several key stakeholders:
  - Government Officials (in this case, the Seattle Public Development Authority)
  - Drivers
  - Companies developing technology to improve car and road safety

Data

- The dataset contained car accidents occurring from 2004-2020 in Seattle, Washington
- Challenges with the dataset: variation in frequency of predictor variables and empty values
- 'SEVERITYCODE' was the predictor variable (Property Damage vs. Injury), with 'INATTENTIONID', 'UNDERINFL', 'LIGHTCOND' 'ROADCOND' and 'WEATHERCOND' used as response variables
- Encoding was changed from the original dataset
- After the data cleaning phase, it was discovered data was unbalanced and skewed towards Property Damage
- The SMOTE package was used from the imblearn library to create an equally proportioned distribution of target variables

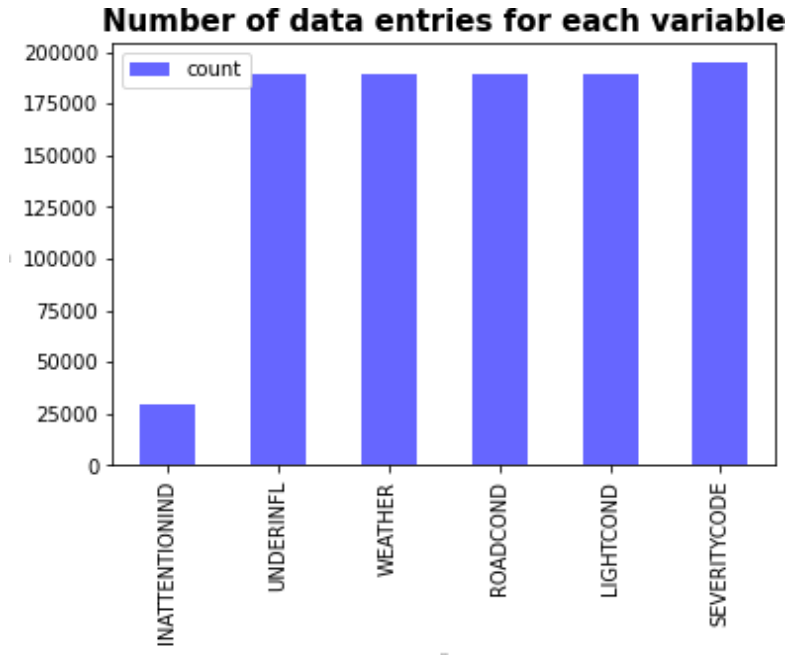


Fig 1: Number of data entries for each variable
To solve the issue of uneven frequency of variables, arrays were created for each column and encoded according to the initial column data in equal proportions

Machine Learning Models and Results

- Logistic Regression and Decision Tree Analysis were used
- Decision Tree: entropy with a max depth of 6 was used for the classifier. After applying SMOTE, the new balanced data was utilized to fit and predict the Decision Tree
- Logistic Regression: the regularization strength was 0.01 and solver was liblinear. The post-SMOTE balanced data was used to predict and fit the Logistic Regression Classifier

Fig 2: Decision Tree Results and Confusion Matrix

	Precision	Recall	F1 Score
0	0.66	0.72	0.69
1	0.41	0.34	0.37
Accuracy			0.58
Macro Avg	0.53	0.53	0.53
Weighted Avg	0.57	0.58	0.57

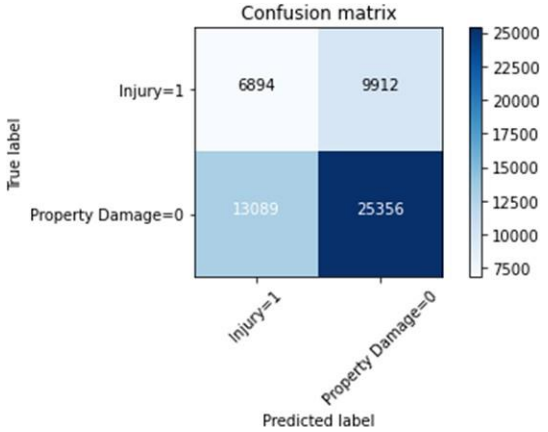
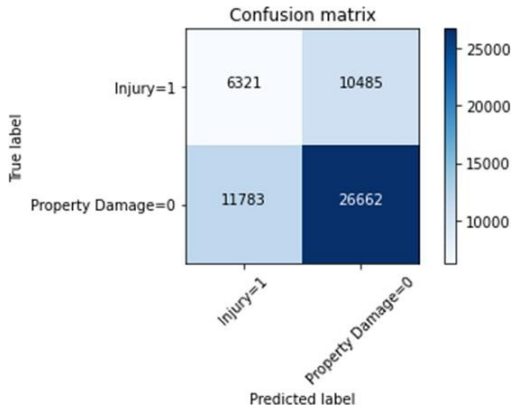


Fig 3: Logistic Regression Results and Confusion Matrix

	Precision	Recall	F1 Score
0	0.72	0.69	0.71
1	0.35	0.38	0.36
Accuracy			0.60
Macro Avg	0.53	0.53	0.53
Weighted Avg	0.61	0.60	0.60



Machine Learning Models and Conclusion

Predictive Model	Avg F1 Score	Property Damage (0) vs. Injury (1)	Precision	
Decision Tree	0.53	0	0.66	0.72
		1	0.41	0.34
Logistic Regression	0.54	0	0.72	0.69
		1	0.35	0.38

- An average F1 score was calculated from Property Damage and Injury
- When comparing the two models, Logistic Regression has a slightly higher F1 score of 0.54, whereas the Decision Tree algorithm has a score of 0.53
- Precision and recall of Logistic Regression is a bit superior, but the average F1 scores are quite close
- In terms of Precision, the best model seems to be Decision Tree while Logistic Regression has more of an unbalanced percentage
- In terms of Recall, Logistic Regression is the more balanced model for predicting the target variable
- It is advantageous to use both models in conjunction with each other
- Accuracy and performance of the model could also be improved with a more balanced dataset and fewer blank values

Recommendations

- The public officials in Seattle can utilize the data to undertake development projects in areas where accidents frequently occur and assess their severity
- The majority of accidents took place on either a block or intersection
- Most accidents occurred under poor lighting, weather and/or visibility conditions
- Both drivers and government officials can make use of the predictive model for accidents occurring based on weather and road conditions
- Companies developing technology to improve road safety can find solutions to minimize future injury and property damage from car accidents based on the predictive algorithm