Accident Severity Probability Prediction in Seattle City

Abednego Kristanto

Coursera IBM Applied Data Science Capstone Project

8 September 2020

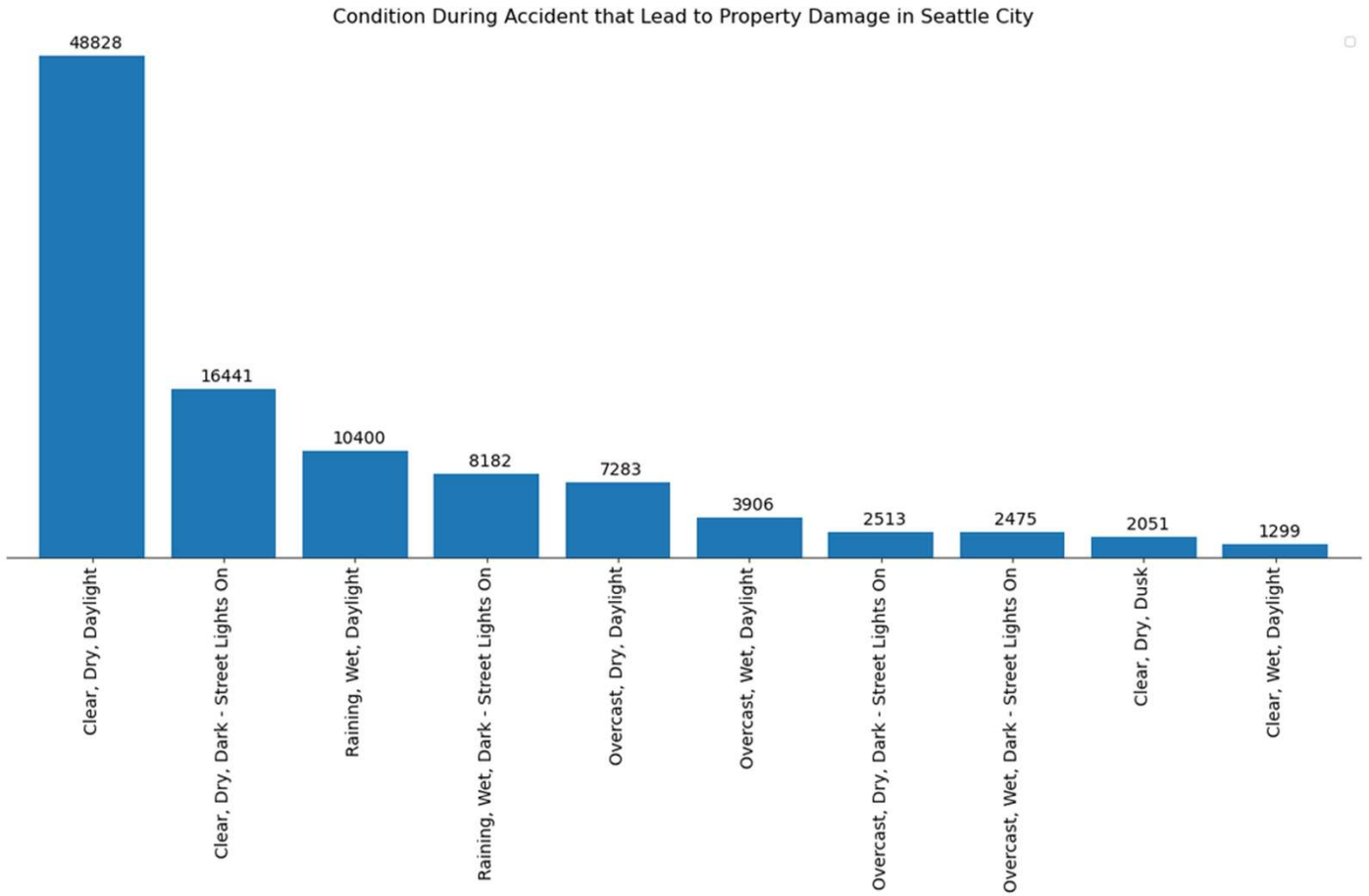
Introduction

- Safety is always everyone priority, however as stated in a well-known Murphy's Law: "Anything that can go wrong will go wrong".
- Therefore a model to predict the traffic accident severity will be useful to improve the driving safety.
- This model will be useful for the government to improve the safety on the road infrastructure, or GPS system manufacturer to give their product more attractive selling point.
- Data that represent road condition that lead to accidents is needed to build such model.

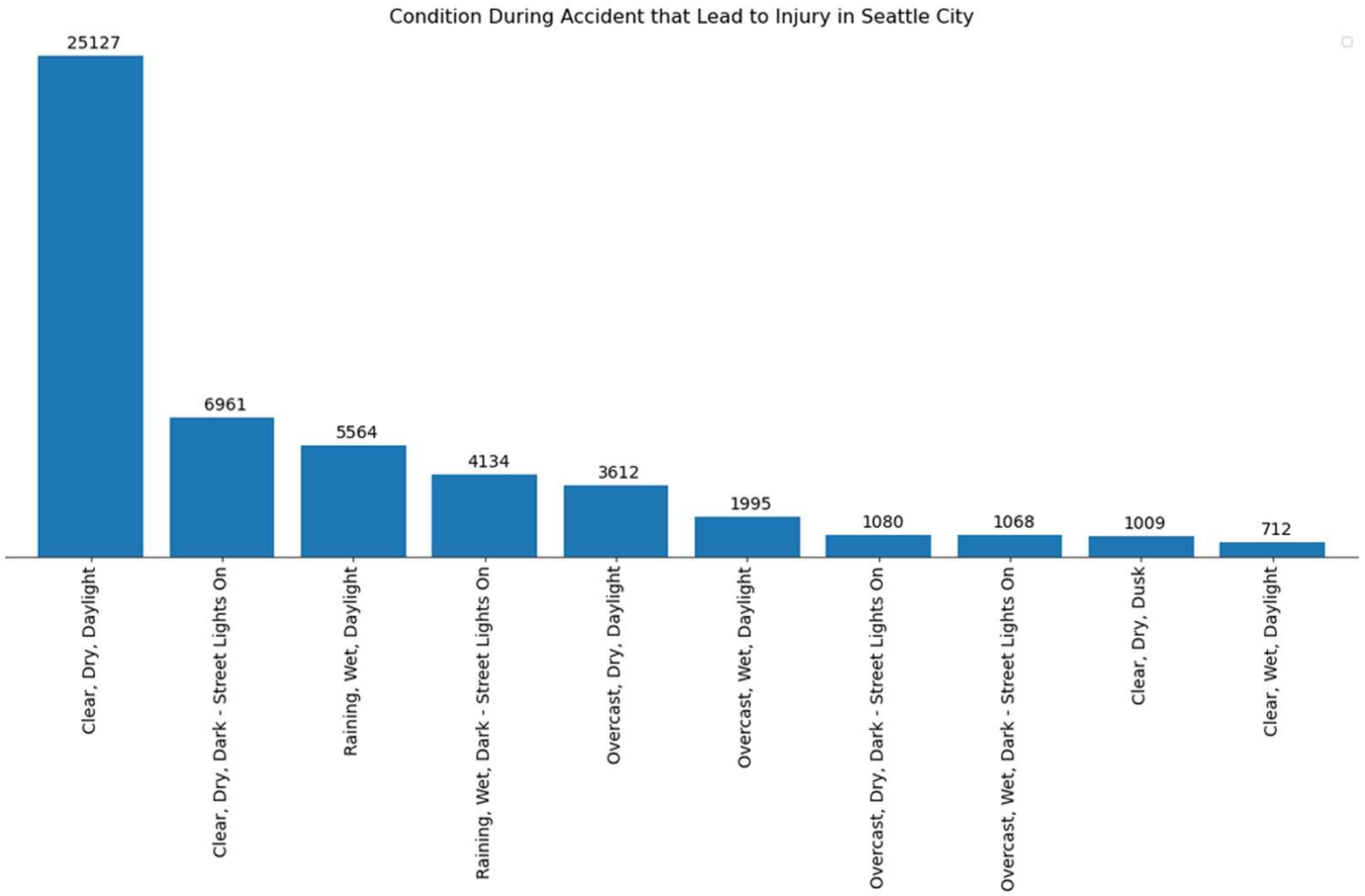
Data Acquisition, Selection, and Cleaning

- Data source: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass /DP0701EN/version-2/Data-Collisions.csv
- Features selected from the data source are location latitude's "X", location's longitude "Y", collision address type "ADDRTYPE", total number of people involved "PERSONCOUNT", total number of vehicle involved "VEHCOUNT", driver's inattention "INATTENTIONIND", driver's under drug or alcohol influence "UNDERINFL", weather "WEATHER", road condition "ROADCOND", and light condition "LIGHTCOND"
- Data cleaning done by dropping and replacing missing value. After this process dataset is containing 189,339 rows of data in 12 columns of features.

Conditions That Lead to Property Damage



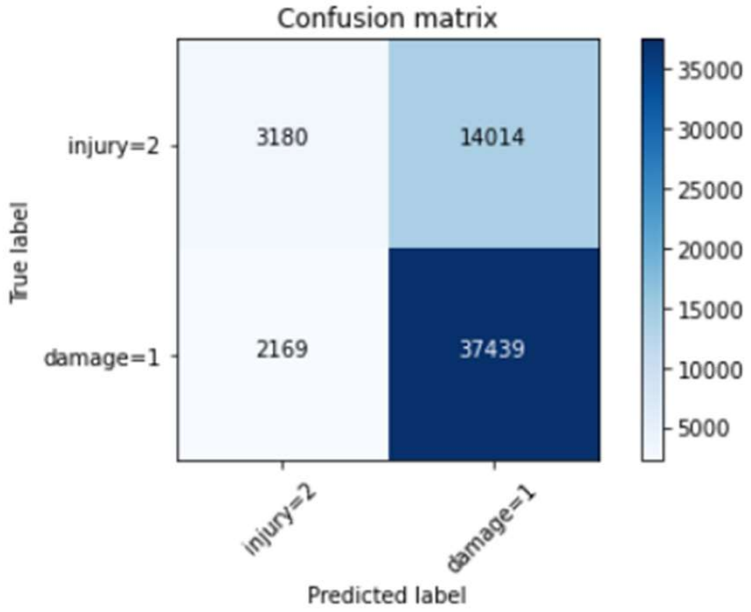
Conditions That Lead to Injury



Logistic Regression Classification Model

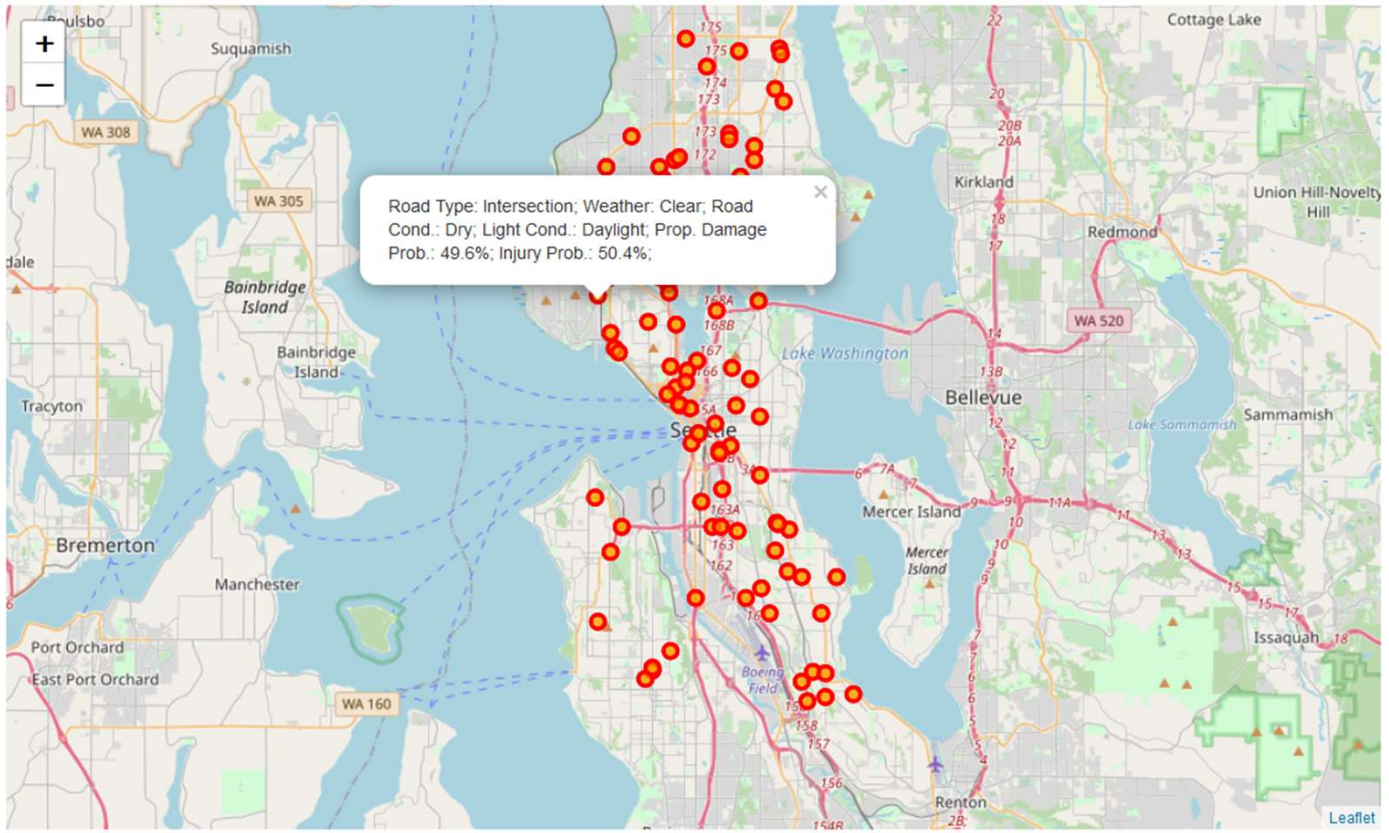
- Categorical data is converted into numerical value with Ordinal Encoder function.
- Data is split using train_test_split function with test size = 30%
- Data is normalized using StandardScaler function.
- Logistic Regression model has been created with regularization value C equal to 0.01, and using solver "lbfgs".
- The evaluation metrics of the model:
  - Jaccard similarity index is 0.7151
  - f1-score with weighted average is 0.6588
  - Logarithmic loss is 0.5692

Confusion Matrix, and Classification Report



	precision	recall	f1-score	support
1	0.73	0.95	0.82	39608
2	0.59	0.18	0.28	17194
micro avg	0.72	0.72	0.72	56802
macro avg	0.66	0.57	0.55	56802
weighted avg	0.69	0.72	0.66	56802

Application Example: Map Plotting



Possible Future Works

- Improving the model accuracy:
  - Adding more severity classes.
  - Adding more features.
  - Adding more data.
- Real-time accident severity probability prediction to determine high-risk road area:
  - Real-time road condition input, including weather and light condition.
  - Predicting the probable severity of accident on the road where the user is driving.

Conclusion

- A model to predict accident severity probability prediction in Seattle City has been created using Logistic Regression Classification model, and produce acceptable accuracy.
- An example application of the model has also been created by plotting testing data features and accident severity probability prediction in a interactive folium map.
- The model is very potential to be developed further in the future, for example by improving its accuracy, and expand its application using real-time data.