Can weather and road conditions predict increased collision severity?

Joel Anderson

The Question

- There are many factors that may increase the risk of severe traffic collisions, such as weather, or road quality
- If information such as this can be used to predict an increased severity of collision, then this predictive capability may be able to be used to reduce severe collisions
- Parties that may be interested in this information are:
  - Drivers, so that they may know when to use extra caution
  - Vehicle manufacturers, who may way to build additional sensors
  - Traffic control officials who control speed limits, where possible

The Data

- This project will use a data set of traffic collisions recorded by the Seattle Police Department and Traffic Records (the data set provided by the course).
- This data set provides a set of descriptive features of each collision, as well as the environment where each collision occurred, such as weather and road conditions.
- The data is labeled with a number from 0 to 3 indicating the severity of each collision.
- To predict whether a collision that occurs has a high probability of being severe, I use the environmental features of this data set as well as the severity labels to train a machine learning model.

Data Cleaning and Preparation

The data was cleaning in the following stages:

- Remove all features except collision severity that cannot be determined until after a collision has occurred. The columns that remain are
  - Weather
  - Road Condition
  - Light Condition
  - Whether the vehicle was speeding
- The speeding column was then dropped because out of > 190,000 samples, only 10,000 don’t have a null value for this feature
- Rows with Null values or values that don’t contain information (such as “Unknown” or “Other”) were dropped.
- Categorical variables (all of them) were transformed into logical columns
- 20% of the data set was selected at random and reserved for testing. The remaining data was used to train a model.

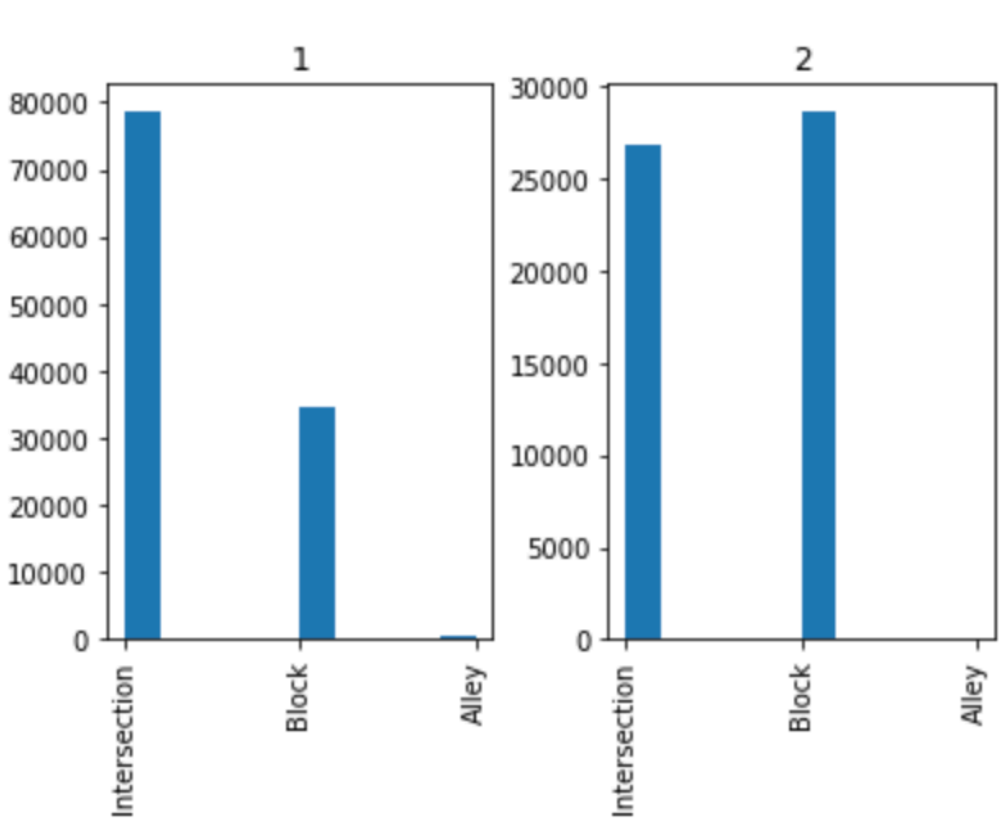
Logistic Regression Model

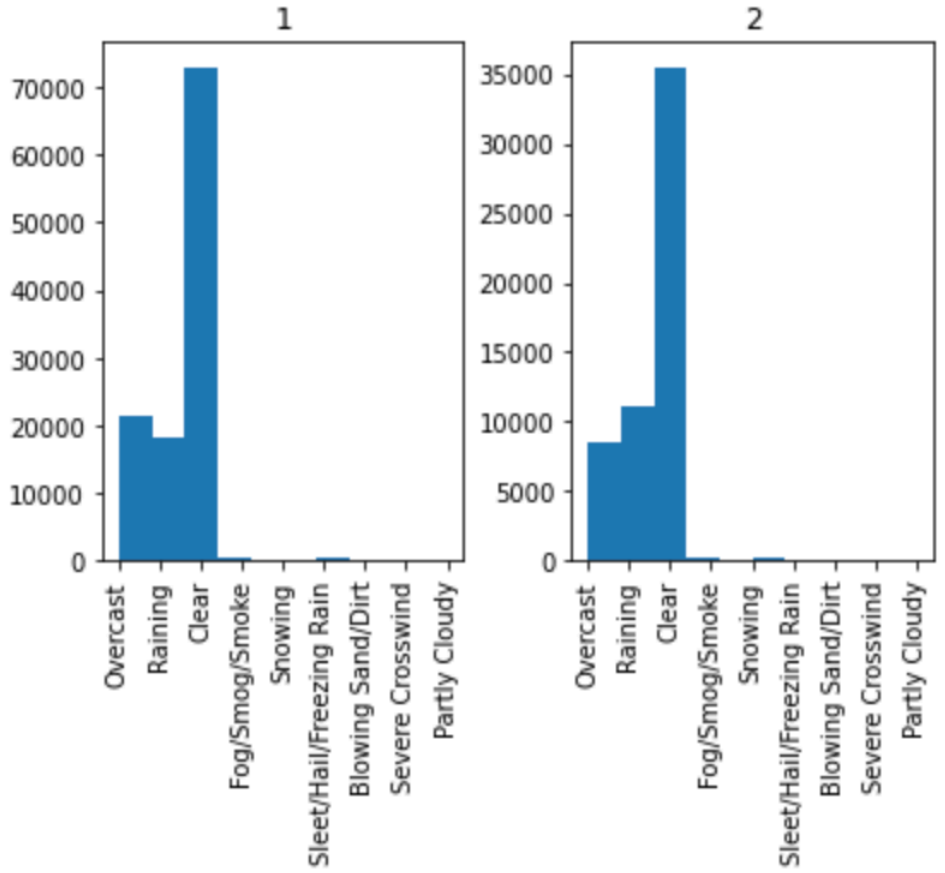
- A logistic regression model was applied to make predictions of accident severity because it can also calculate probabilities for each class.
- Parameter tuning using the test set yielded an optimal inverse regularization parameter of C = 0.066.
- However, even the best performing linear regression model trained only had a Jaccard similarity score of 0.45.
- This model only predicted severe collisions in a single case. Even for this nearly-trivial model, the unbalanced data set allowed this model to achieve a modest Jaccard score.

K-Nearest-Neighbors Model

- Given the lackluster performance of the logistic regression model for this problem, a K-nearest-neighbors model was tried.
- Parameter tuning gave an optimal number of nearest neighbors used, k, of 1. This model achieved a Jaccard similarity score of only 0.23, however this model performed much better when measured by weighted F1 score (0.6 compared to 0.53).
- This model also did not predict less-severe collisions for almost every case.
- However, this is ultimately a rather poor-performing predictor

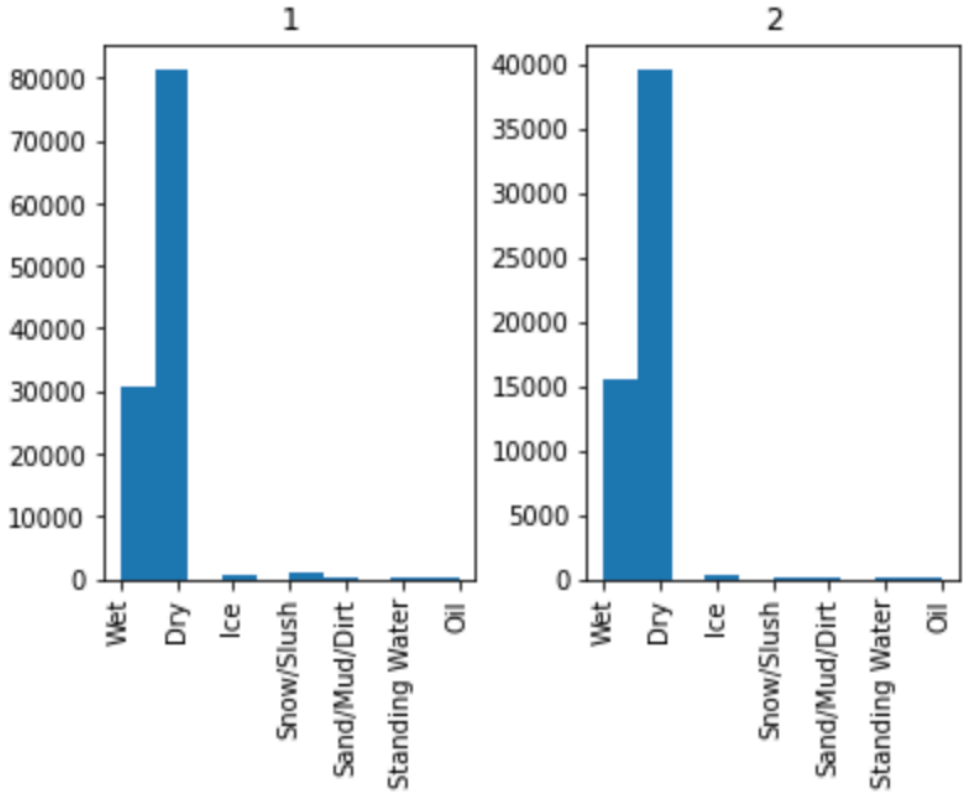
Revisiting the Data

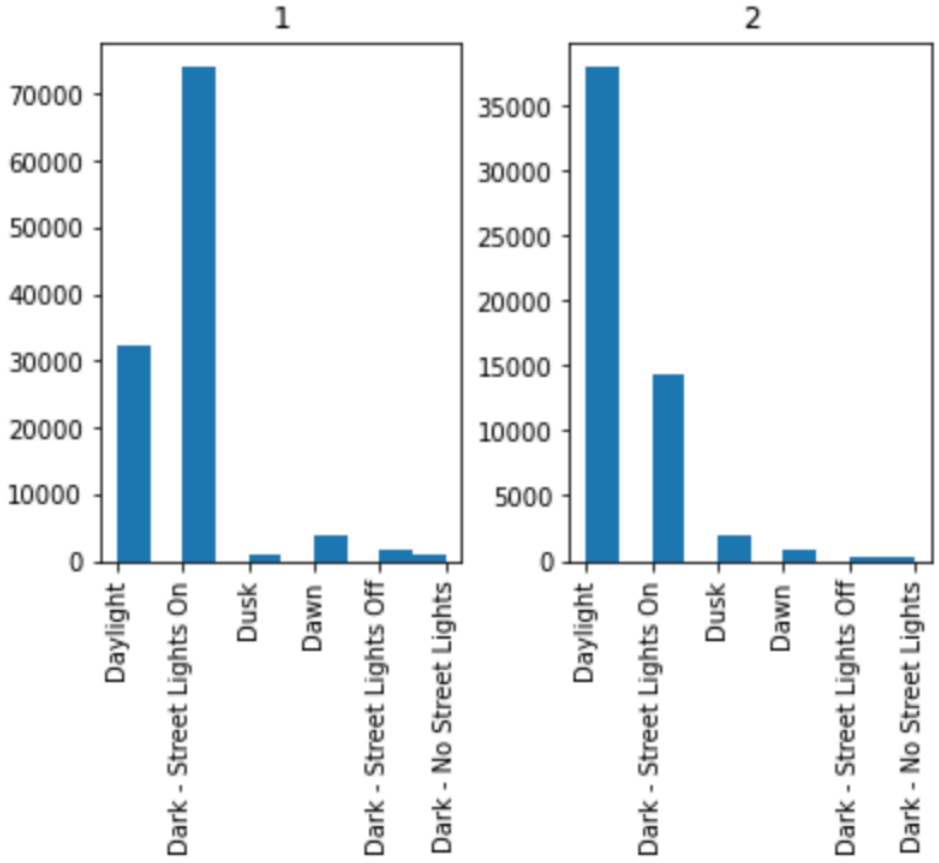




Only minimal differences between distributions of features separated by severity level 1 or 2. Left: Address type. Right: Weather

Revisiting the Data





Only minimal differences between distributions of features separated by severity level 1 or 2 on left: Road condition. On right, light level does seem to show some differences.

Conclusions

- Ideally, more types of features are needed if one wants to predict more severe collisions **before** they happen.
- Though the K-Nearest-Neighbors algorithm performed better than the logistic regression algorithm, neither performed exceptionally well.
- In this case, only a single variable (when taken alone) appeared to be even somewhat predictive of accident severity.