Predicting Car accident severity

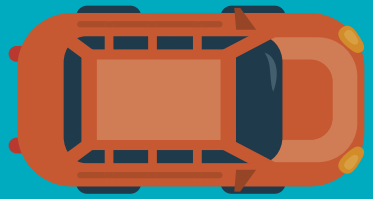
By: David Chabra, Kai-Li Chang, Harry Kim, Chan Wai Wong

The Data

- Our data is a limited selection from the a Countrywide Traffic Accident Dataset (2016 - 2021)
- The unit of observation is a single car accident
- Each row contains information on time of the incident, location of the incident, some details about how it happened, and a description
- We have two data sets one for training which includes Severity as a variable and testing which excludes Severity
- In the raw data only 0.85% of entries are missing values in the training data and 0.91% for testing data

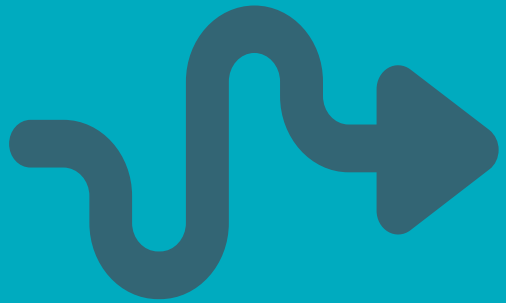
Our Goal: Predict Car accident severity from our data and added predictors

STEP 1



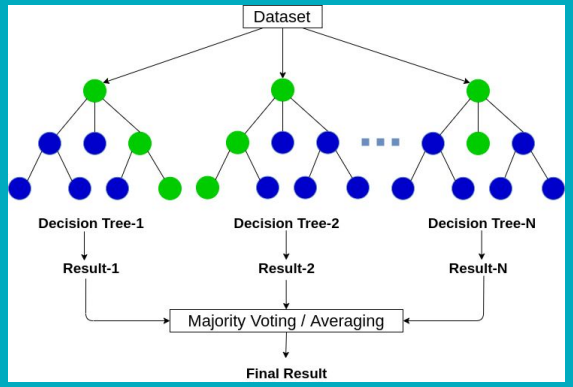
Explore our data to find useful predictors

STEP 2



Mine information, join new data, and remove complex variables

STEP 3



Fit Random Forest Model to make predictions

Exploratory Analysis

- From our expority analysis we concluded we had many strong predictors but most require significant transformation
- Description is a very useful text variables but extremely complex in its original form
- Much of the location information has too many levels as factors
- The time variables are very useful but again should be simplified
- We also noticed many possible external data to join based on zip code, city, and county
- The variable country is useless since every observation is in the U.S.

Transforming Our Data

01

Time variables

Covid Year, Holiday, and Duration

02

Description variable

Dummy variables for important words and phrases obtained from word cloud

03

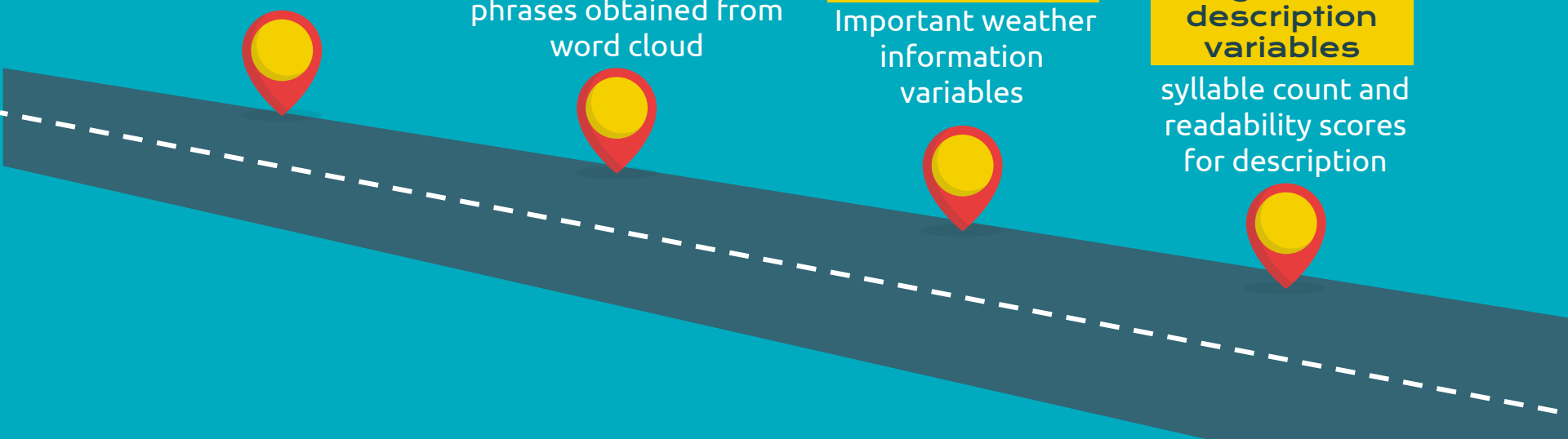
Weather variable

Important weather information variables

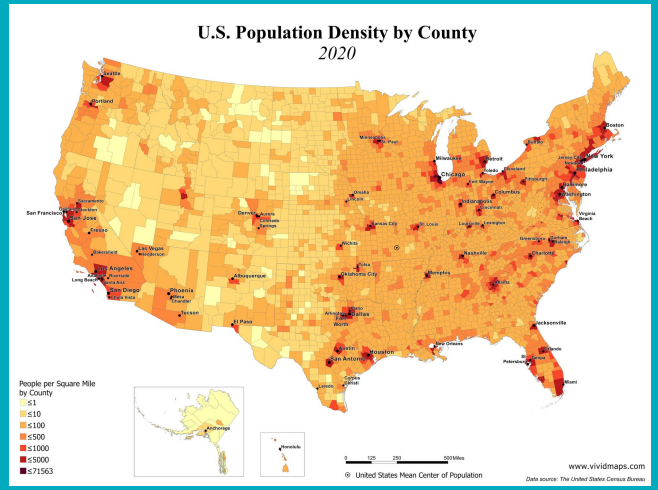
04

Sylcount description variables

syllable count and readability scores for description



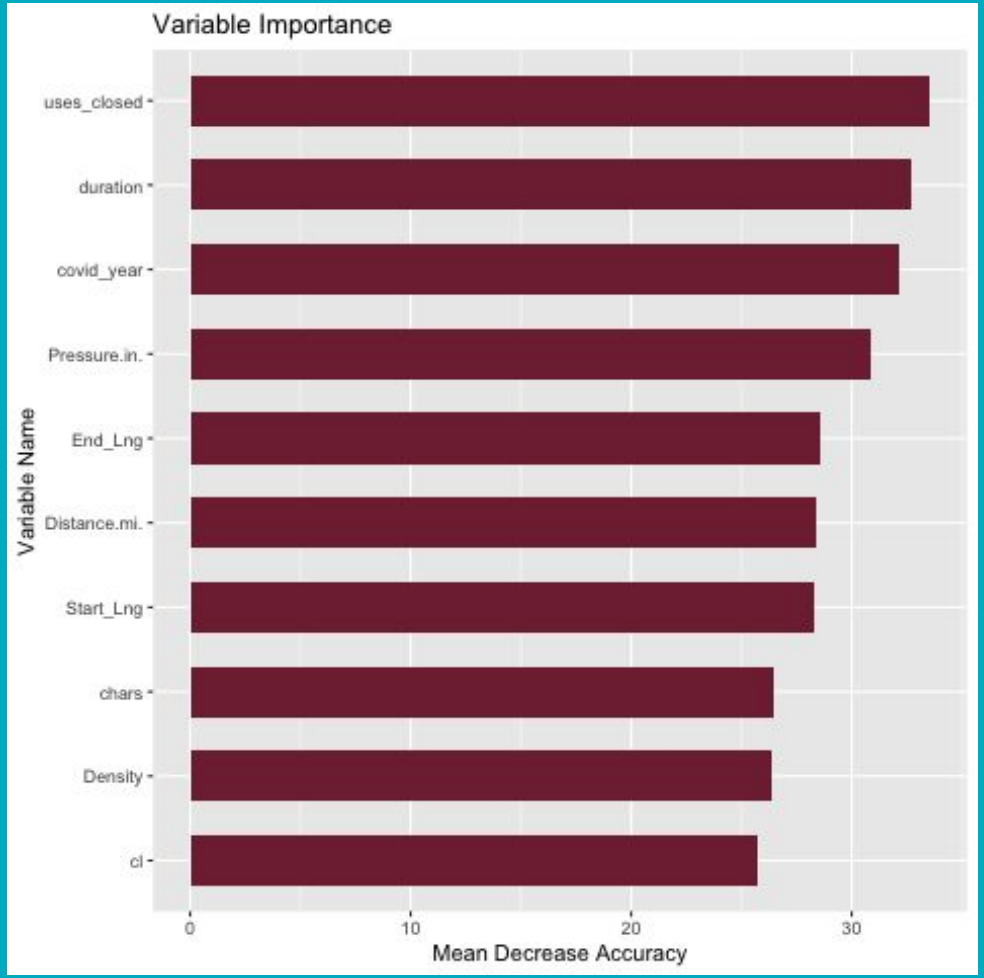
Adding External Data and Removing Complex Variables



- We…
  - Joined data on total population and population density by zip code from the Geographical repository maintained by Opendatasoft.
  - Removed factor variables with many levels like State since random forests tend to overvalue these variables
  - Removed two variables with more than 5% NAs of the data
  - Removed variables we had created that had too low or too high variability to be useful

Our Model (Random Forest)

- We fit a random forest model to our data and found several stand out predictors
- Variables obtained from the description were particularly important
- Our final model accuracy for testing was 94.25% which combines public and private scores



References

- Slide template was provided by Slidesgo
- Text Mining given by https://voyant-tools.org/
- Random Forest graphic on slide 3 was provided by https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/
- US Population Density Visual by https://vividmaps.com/us-population-density/
- Zipcode total pupulation and population density https://data.opendatasoft.com/explore/dataset/georef-united-states-of-america-zc-point%40public/table/