Accident Prediction Models for Illinois Roadway Network

CS 498: DSG - Graduate Project Presentation

Jacob Mathew (jmathew7)

Trisha Das (trishad2)

Jinsong Cui (jinsong4)

Introduction

Motivation:

Each year in Illinois, there are around 144,000 accidents and 7000 (~5%) people lose their lives

Typically locations with high number of accidents have a systemic cause for their high numbers

Having the ability to predict the number of accidents and identify locations likely to have a high accident count can enable prioritization of locations for safety improvements.

Goal:

Develop a prediction model to estimate the likelihood of crashes as a function of the features of a location using historical crash data

Data

Obtained real data from Highway Safety Information System (HSISinfo.org)

Data obtained for years 2006-2010 (5 years)

- Accident data (contains information about characteristics of accident)
- Roadway data (contains information about characteristics of roadway)
- Vehicle data (used to remove pedestrian accidents from analysis)

Before start of analysis the data was:

- Merged so that each accident is corresponding to a segment of the road.
- Filtered data to have a meaningful dataset
- Missing values were imputed (speed limit)
- Based on background knowledge, categorical features with high number of categories were reduced

Data Filtering

To filter data:

1. Remove pedestrian accidents
2. Remove locations where number of lanes were marked as 0
3. Remove locations with traffic volume (aadt) = 0
4. Remove locations where lane width (lanewid) < 10 feet
    a. The standard is 12 feet, but removing locations with lanewid < 12 feet removes 35000 accidents
    b. Removing locations with lanewid < 10 removes only 1700 accidents

Improving Features

1. Curve Radius (curv_rad): Gives the radius of curve in feet

   Changed to binary categorical variable

   - If curv_rad = 0 => straight segment
   - If curv_rad > 0 => curved segment

2. Median Type (med_type): One of the eight type of medians

   Changed to binary categorical variable

   - No median
   - Median Present

Improving Features

3. Surface Type (surf_typ): One of the twenty five type of surface

    Changed to categorical variable

        Flexible surface (i.e. asphalt pavement)

        Rigid surface (i.e. concrete pavement)

        Others

Imputing missing values

Speed limit: Posted roadway speed limit

There were 105,209 (12.7%) missing values

Speed limit was available in multiples of 5, and we divided it into 11 classes

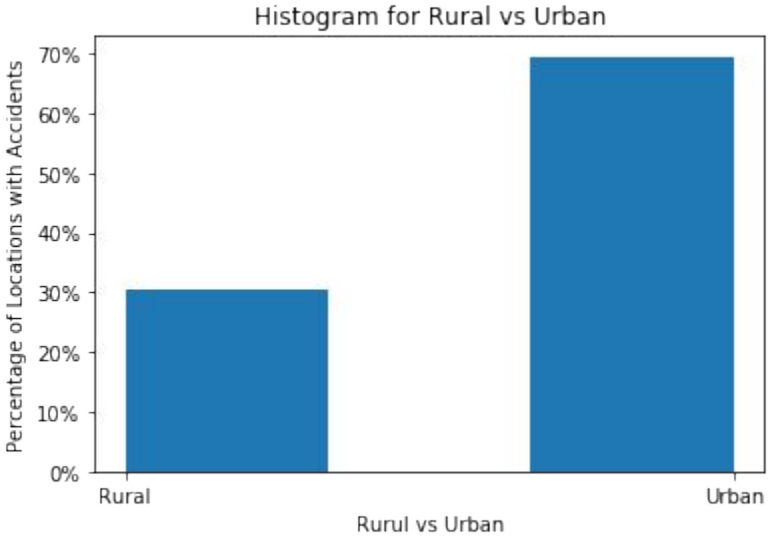
The missing values were imputed based on multi-class classification Random Forest model.

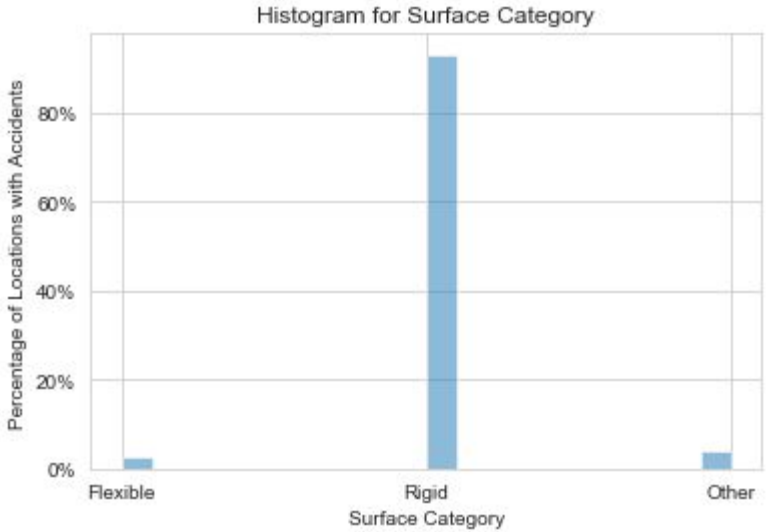
Accuracy of speed limit model is 87% on test data

Features used for prediction

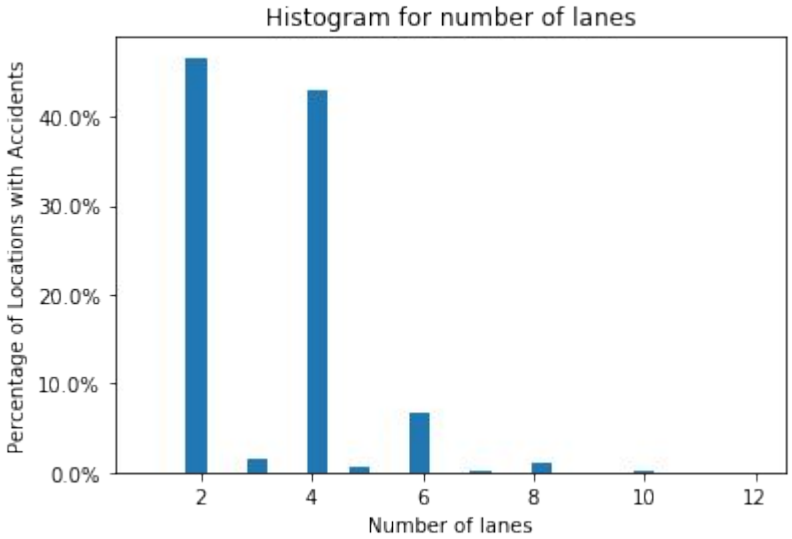
1. Number of lanes
2. Lane width
3. Curve?
4. Length of Segment
5. Access control at the location
6. One way?
7. Median?
8. Speed Limit
9. Surface Type
10. Rural or Urban
11. Roadway classification

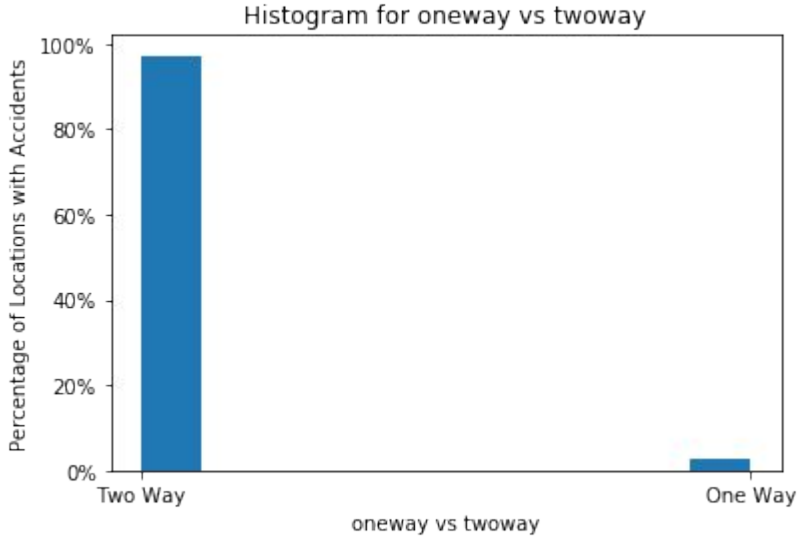
Data Distribution



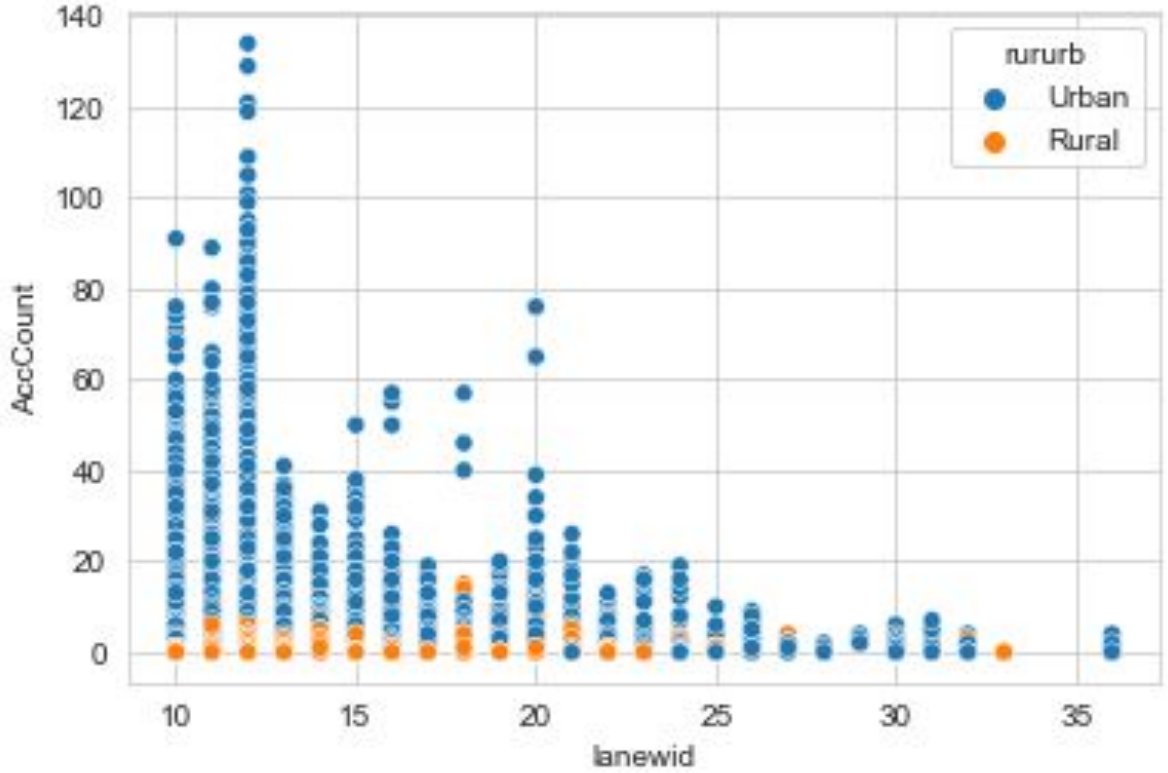


Data Distribution





Lane-width vs Accident Count



EDA Summary

Identify the continuous variables and discrete variables and apply proper data cleaning techniques.

Identify the correlation between each column variable and visually inspect outliers by pair plot

Hypothesis:

The accident count is a variable (observation) based on many variables in the dataset like road types and conditions (hidden state) but not all variables.

The accidents didn’t happen at the same time, so can be treated as a sequenced data.

Methods

Accident count is positive integer

Preliminary analysis used a binary classification model

On further analysis we classified locations with accidents as

- Zero accident location
- Low accident location (1 - 30 accidents)
- Medium accident location (31- 50 accidents)
- High accident location (> 50 accidents)

Models Used

Classification Model

1. Logistic Regression
2. XGBoost
3. Neural Networks

“Hidden Markov Model”

Used to predict the sequence of occurrence of accidents.

Classification Model Pipeline

Train data: 827246 data points

Test data: 2000 data points

Data normalized using standard scalar

Models validated using k-fold cross validation

Motivation for HMM

We have data available as a sequence.

A location which witnessed high number of accidents could experience a reduction in the number of accidents in the following year as location may be improved or vice versa

We assumed two states that the system could be in

1. High likelihood of accident
2. Low likelihood of accident

Emissions (or observations) is the accident count observed

Divided into 4 categories: ‘Zero’, ‘Low’, ‘Medium’, ‘High’

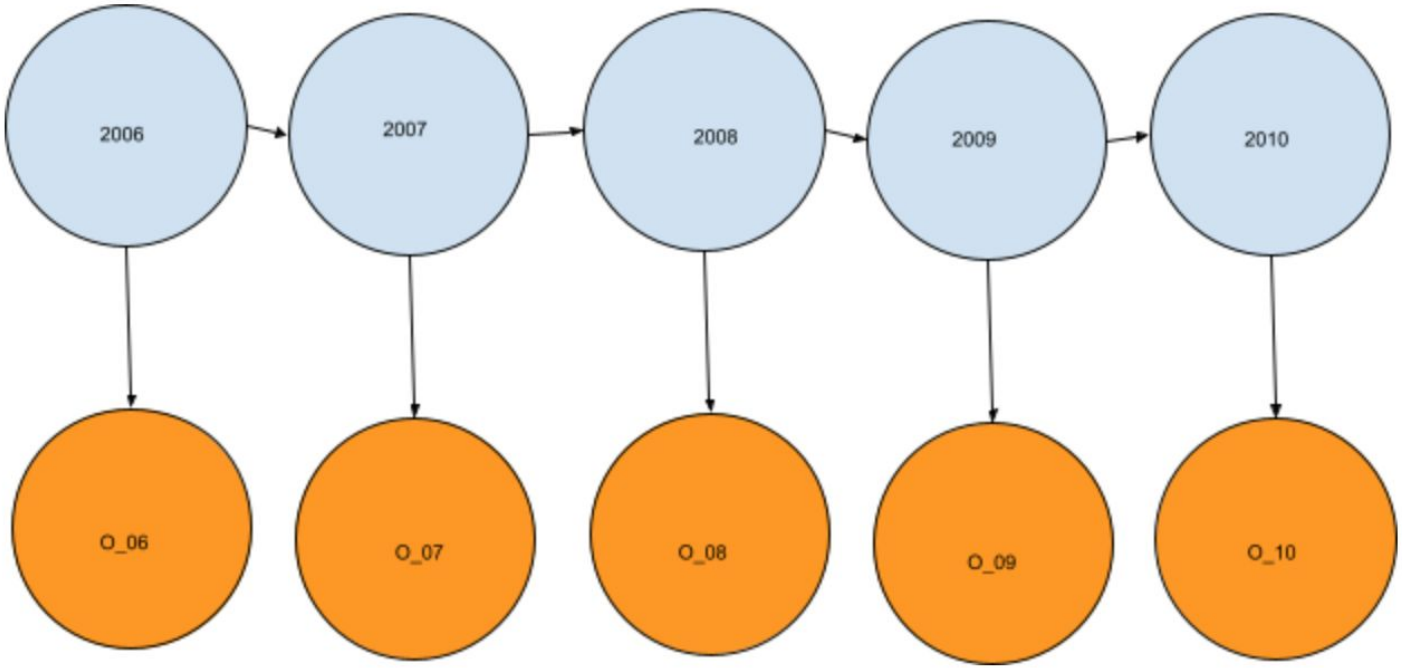
Transmission and Emission Probability

As an observer,

- It is difficult to say which state (high likelihood or low likelihood) a location is in in any particular year
- Without knowing the states of the previous observations, it is difficult to estimate the emission probabilities

Therefore

- We reduced the problem to one ‘hidden state’
- With four possible values for the observations as previously listed



O_i = Observation for year i

P(O_10|O_06, O_07, O_08, O_09) = ?

‘HMM’ cont.

Using data for the years 2006-2009, we estimated emission probabilities

Used multi-class logistic regression to estimate emission probabilities

Emission probabilities depend on the characteristics of the location

Test dataset is the data for the year 2010, which had corresponding values in the previous 4 years.

Discussion of Results

The models developed were compared based on the accuracy metric

Model	Accuracy
Logistic Regression	68.6%
XGBoost	69.85%
Neural Networks	73.60%
“Hidden Markov Model”	79.15%

Conclusions and Recommendations

The study developed a predictive model to estimate the likelihood of crashes on highway network in Illinois using real world data

Accident count is positive integer value. This was modified into categories to use classification algorithms to build predictive models

Missing values in data was imputed

Data cleaned and filtered and modified before analysis

Based on study, the Hidden Markov Model gives the best predictive accuracy and therefore is recommended.

Limitations and Future Work

1. Limitations in data:
    a. The available data for the years 2009 and 2010 are reduced due to higher restrictions
2. Classification models doesn’t consider accident history of a location while making prediction
3. The transmission probabilities in the HMM might change if the characteristics of a location changes. This is not considered in our study.

Future work

1. Model could be extended to predict the severity of a crash along with the likelihood of crash

Thank you

Questions/ Comments/ Feedback?

Jacob Mathew(jmathew7)
Trisha Das(trishad2)
Jinsong Cui(jinsong4)