Accident Prediction Models for Illinois Highways: Progress Report 1

Jacob Mathew (jmathew7)

Trisha Das (trishad2)

Jinsong Cui (jinsong4)

## **Introduction**

This report outlines our progress we have made towards the project. We have obtained data from Highway Safety Information System which supports the work. A detailed inspection of the data was done to convert into a meaningful and usable form. Based on the filtered data, we performed EDA and developed preliminary models to estimate accident probability. The rest of this report discusses our progress so far.

## **Methods**

This section discusses the data merging, cleaning, EDA and the preliminary models that were used in the analysis.

### **Data Merging**

The data used in this project was obtained from the Highway Safety Information System ([hsisinfo.org](http://hsisinfo.org)). The collected data came in four files: roadway file, accident file, vehicle file, and occupancy file. Details about these files were provided in the appendix of the project proposal. Among these four files, three are primarily important for this project: the roadway file, the accident file and the vehicle file.

From the accident file, the records that involved a pedestrian accident were removed. This was obtained by comparing the *“case\_no”* field in the accident file to the corresponding *“case\_no”* field in the vehicle file and examining the *“veh\_type”* field in the vehicle file. *“veh\_type”* value of 98 corresponds to accidents involving pedestrians.

In order to merge the roadway file and the filtered accident file, the *“cnty\_rte”* field in the roadway file was compared to the *“cntyrte”* in the accident file, and the *“begmp”* (Beginning Mile Post) and *“endmp”* (End Mile Post) in the roadway file were compared to the *“milepost”* field in the accident file. Each accident was then assigned to a corresponding record in the roadway file adding a new column called *“AccCount”* which gives the number of accidents observed in that segment in that year. The variables *“severity”* and *“weather”* from the accident file were merged to the roadway file in a similar way.

### **Data Cleaning**

The data was passed through a few filters to ensure that the data is meaningful. The filters include:

1. Remove records where the number of lanes (*“no\_lanes”*) is 0

2. Remove records where the lane width (*“lanewid”*) is < 10 feet (The standard lane width for a road segment in the USA is 12 feet. However, using this filter removes significant number of records and accidents and therefore, we chose to remove only records with lane width <10.)

3. Remove records with annual average daily traffic (*“aadt”*) as 0

4. Remove the column traffic control (*“trfc\_cntrl”*) as more than 60% of the records in this field are NA

After data cleaning, there were 881,746 accidents over the period 2006-2010 in Illinois.

### **Exploratory Data Analysis (EDA)**

With pair plot, we are able to identify the correlation between each column variable and visually inspect outliers. In our dataset, “*curv\_rad”* , “*medwid”* are the main variables which have the most outliers. When doing the data cleaning, these variables need to be done more carefully. More effort should be put on variables with large variance like *“begmp”*  and *“endmp”* with *“seg\_lng”.* For discrete variables like *“access”,* we should build models suitable for both continuous and discrete variables or put discrete variables apart and build a separate model to reduce computation cost.

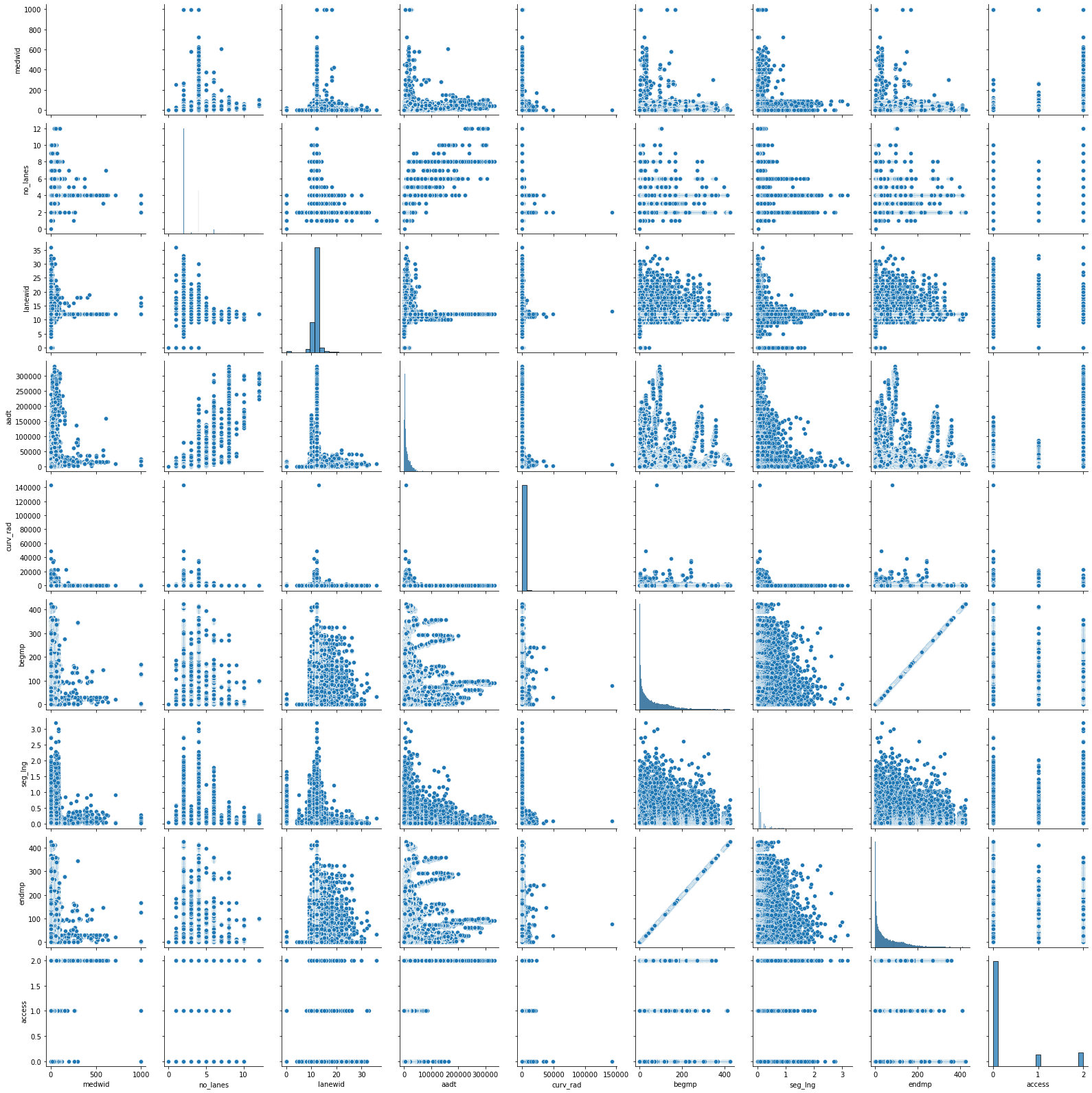


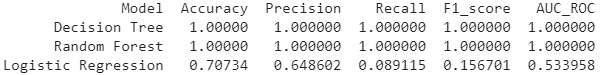
Figure 1: Pair plot of filtered data, for visual inspection and interpretation.

### **Preliminary Models**

We used logistic regression, decision trees and random forest models as our preliminary models. We split the 5-years data into training and test sets (25% test set) and trained the models on the training set. Then we used the test data for prediction and performance scoring. The *“AccCount”* variable was treated as a binary variable (i.e. 0 if no accidents were observed and 1 if at least one accident was observed) and a new variable named ‘*Accident*’ was created to contain this binary information while keeping the ‘*AccCount*’ as it is.

## **Results**

Here is the summary of the model scores of the three models we used.



## **Current Challenges**

Some of the challenges we identified in our work are:

1. Accident count is not a binary variable. The non-binary nature of the dependent variable is to be incorporated into our models.
2. The posted speed limit variable is unavailable in a lot of the cases. However, this variable is intuitively important to predict accidents. One challenge is to develop a model to fill the missing/unknown values in this column.
3. In the years 2009 and 2010, only accidents that cross a monetary threshold that is higher than the threshold for the previous years are reported. Therefore, the number of accidents in 2009 and 2010 are lower than the number of accidents in the previous years. This needs to be incorporated
4. The accident count data is imbalanced. There are a lot more locations with no accident than locations that witnessed an accident. This should be considered in the model.
5. The team currently doesn’t have geographical data as to where the road locations are situated. This is very helpful to make a meaningful display of our results.

## **Plans for Upcoming Weeks**

A few items we were planning to explore in the upcoming weeks include:

1. Feature engineering
2. Feature selection
3. Categorize *‘AccCount’* into multiple classes. For example: 0-5 accidents: low , 6-15 accidents: medium, > 16 accidents: high.
4. Build Neural Networks for multiclass-classification
5. Build HMM model if other models classify poorly