CS584 – MACHINE LEARNING

FALL 2016

Predicting occurrence of major accident

Group Members: Hariharan Devarajan

Table of Contents

Predicting occurrence of a major accident

Group Members: Hariharan Devarajan

Task

The task is to create a model which could predict whether major accident would occur. Here major accident is defined as more than one vehicle colliding. The motivation behind this task is the fact that accidents in Illinois state are generally major and these are only increasing every year. So analyze and come up with a model which could predict and hence could describe patterns which help reduce accidents by proactive measures. The Problem is a classification problem to identify whether major accident would occur or not. The fact that it will need to analyze both conditions of roads and also driving make this an interesting problem.

Dataset

I have got the crash data of 2014 year. It has a lot of columns like place where accident occur. The conditions and so on. The dataset has 292019 records with 80 features. Here is a summary of each feature and the target variable.

- Features
  - OBJECTID : Identifier of each record
  - geodb_oid : Geographical db id
  - ROUTE : Route in the state
  - Case_Id : Case Identifier
  - YEAR : Year when crash happened
  - MONTH : month when crash happened
  - DAY : Day when crash happened
  - HOUR : Hour when crash happened
  - DAY_O_WEEK : Day of the week when crash happened
  - INJURIES : No of injuries occurred
  - FATALITIES : No of fatalities from the incident
  - COLL_TYPE : Type of Collision
  - WEATHER : Whether Type
  - LIGHTING : Type of Lightening
  - SURF_COND : Condition of the surface
  - RD_DEFECT : Road Defects
  - RD_FEATURE : Road Feature
  - TRAF_CNTRL : Traffic Control was present or not
  - COUNTY : Country where this happened
  - TOWNSHIP : Township where this occurred
  - TS_ROUTE : Route info
  - MILE : How many miles the accident was spread
  - CITY : City where this happened
  - DRIVER_1 : Condition of Driver
  - VEH1_TYPE : Type of Vehicle 1 involved in accident

- VEH1_SPECL : Vehicle ownership
- VEH1_DIR : Vehicle was heading which direction
- VEH1_MANUV : Type of maneuver performed by vehicle
- VEH1_EVNT1 : Exact event with vehicle 1
- VEH1_LOC1 : Location where vehicle 1 was found
- VEH1_EVNT2 : Second event with vehicle 1
- VEH1_LOC2 : Second location where vehicle 1 was found
- VEH1_EVNT3 : Third event with vehicle 1
- VEH1_LOC3 : Third location where vehicle 1 was found
- DRIVER_2 : Condition of Driver
- VEH2_TYPE : Type of Vehicle 2 involved in accident
- VEH2_SPECL : Vehicle ownership
- VEH2_DIR : Vehicle was heading which direction
- VEH2_MANUV : Type of maneuver performed by vehicle
- VEH2_EVNT1 : Exact event with vehicle 2
- VEH2_LOC1 : Location where vehicle 2 was found
- VEH2_EVNT2 : Second event with vehicle 2
- VEH2_LOC2 : Second location where vehicle 2 was found
- VEH2_EVNT3 : Third event with vehicle 2
- VEH2_LOC3 : Third location where vehicle 1 was found
- DRIVER_3 : Condition of Driver
- VEH3_TYPE : Type of Vehicle 3 involved in accident
- VEH3_SPECL : Vehicle ownership
- VEH3_DIR : Vehicle was heading which direction
- VEH3_MANUV : Type of maneuver performed by vehicle
- VEH3_EVNT1 : Exact event with vehicle 3
- VEH3_LOC1 : Location where vehicle 3 was found
- VEH3_EVNT2 : Second event with vehicle 3
- VEH3_LOC2 : Second location where vehicle 3 was found
- VEH3_EVNT3 : Third event with vehicle 3
- VEH3_LOC3 : Third location where vehicle 3 was found
- DRIVER_4 : Condition of Driver
- VEH4_TYPE : Type of Vehicle 4 involved in accident
- VEH4_SPECL : Vehicle ownership
- VEH4_DIR : Vehicle was heading which direction
- VEH4_MANUV : Type of maneuver performed by vehicle
- VEH4_EVNT1 : Exact event with vehicle 4
- VEH4_LOC1 : Location where vehicle 4 was found
- VEH4_EVNT2 : Second event with vehicle 4
- VEH4_LOC2 : Second location where vehicle 4 was found
- VEH4_EVNT3 : Third event with vehicle 4
- VEH4_LOC3 : Third location where vehicle 4 was found
- DUP_CD : No Idea
- REC_TYPE : Type of Injury

- XCOORD : X Coordinate where accident occur
- YCOORD : Y Coordinate where accident occur
- INTERSEC : Was it an intersection
- SFE : Vehicle Number
- AGENCY_NUM : Agency Identifier
- RUNDATE : Date when it was run
- WorkZone : Was are a workzone
- WorkZoneTy : Work Zone type
- WorkersPre : Were workers present
- ExceedSpee : Was speed exceeded
- CellPhoneU : Was cell phone used.
- NUM_VEH : This is the Target variable. Number of vehicle(s) involved

Data source

The Data source was IDOT intern. I has asked him to check with his manager to get some data on crash. As the data shared was not sensitive they were ready to share the data.

Target variable

The number of vehicles involved in Crash (“NUM_VEH”)

Features

There are 80 features and I have described then in the dataset region

Data size

I have 292019 instances

Preprocessing

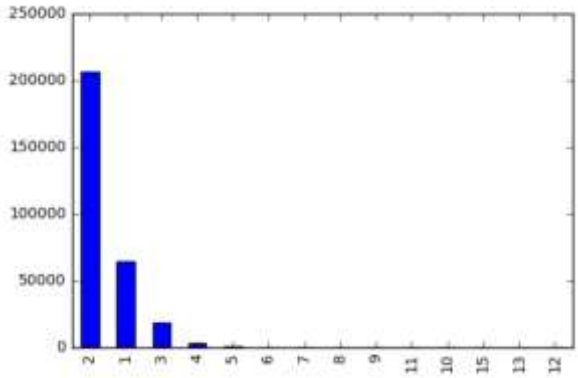
I have done the following preprocessing on the data:

1. Filled mean in the NA columns of numerical data.
2. Removed the columns which were effects of accident not cause like injuries, fatalities, etc.
3. I removed all the Identifier columns.
4. I converted categorical data to numerical data.
5. I converted NUM_VEH my target to binary by rule 1 if >1 else 0

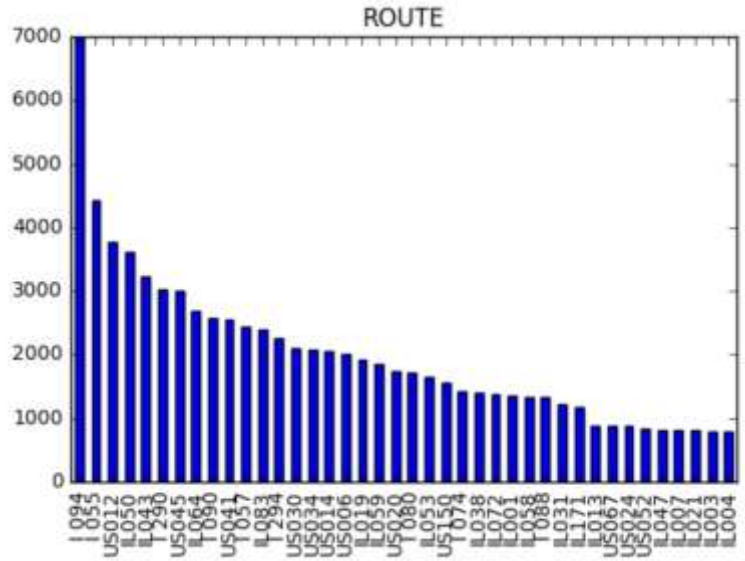
Visualization

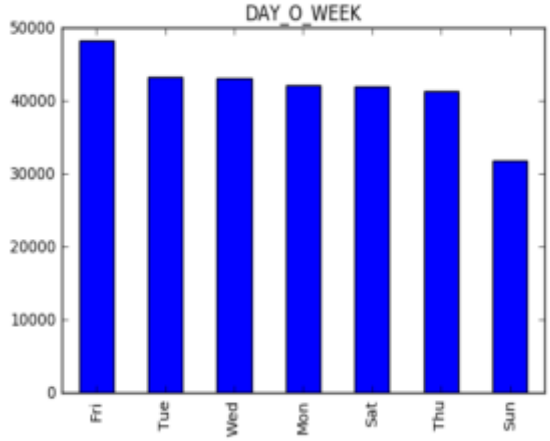
Target

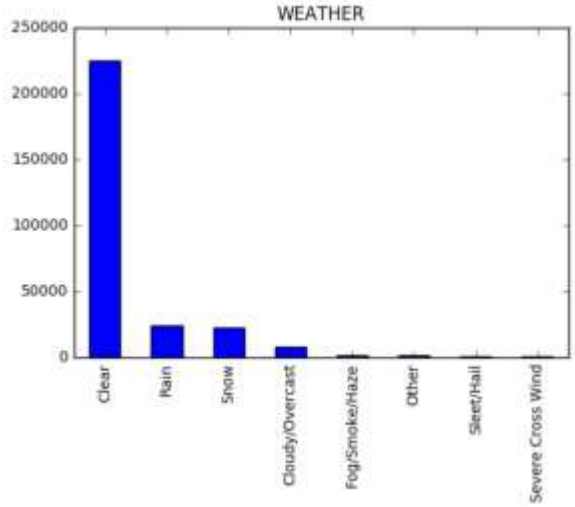
```
count    292019.000000
mean          1.869991
std           0.580815
min           1.000000
25%           2.000000
50%           2.000000
75%           2.000000
max          15.000000
```

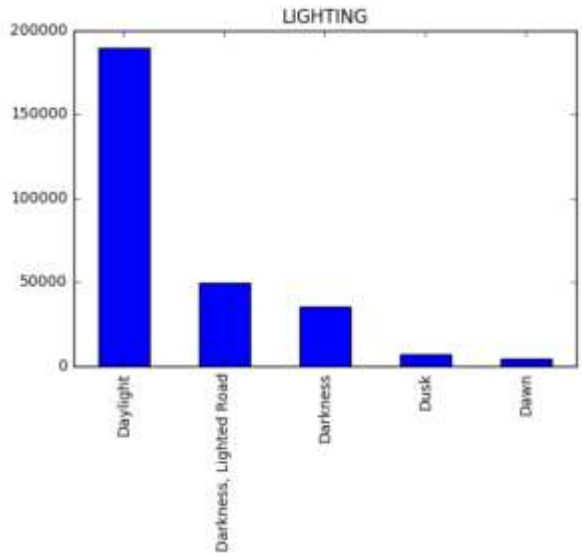


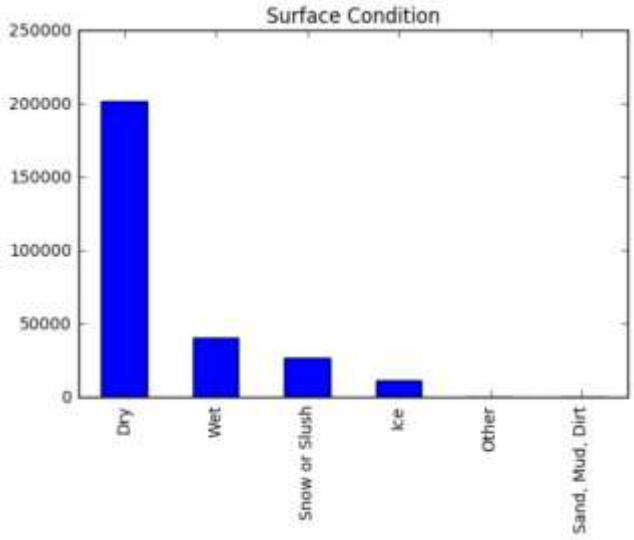
Features

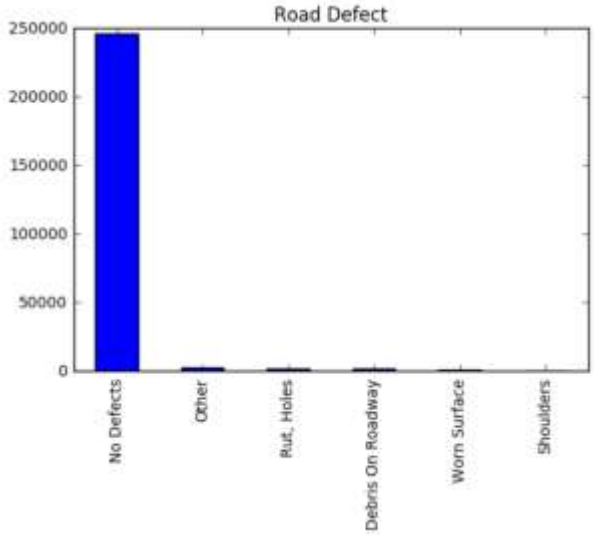

ROUTE
7000
6000
5000
4000
3000
2000
1000
0
I094
I055
US012
IL050
IL043
I290
US045
IL064
I090
US041
I057
I083
I294
US030
US034
US014
US006
IL019
IL059
US020
I080
IL053
US150
I074
IL038
IL072
IL001
I088
IL031
IL171
IL013
US067
US024
US052
IL047
IL021
IL003
IL004

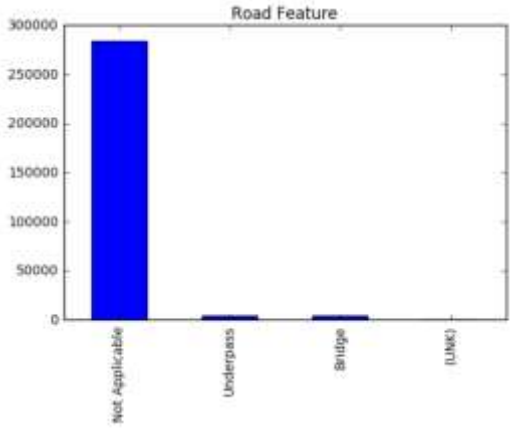

DAY_O_WEEK
50000
40000
30000
20000
10000
0
Fri
Tue
Wed
Mon
Sat
Thu
Sun

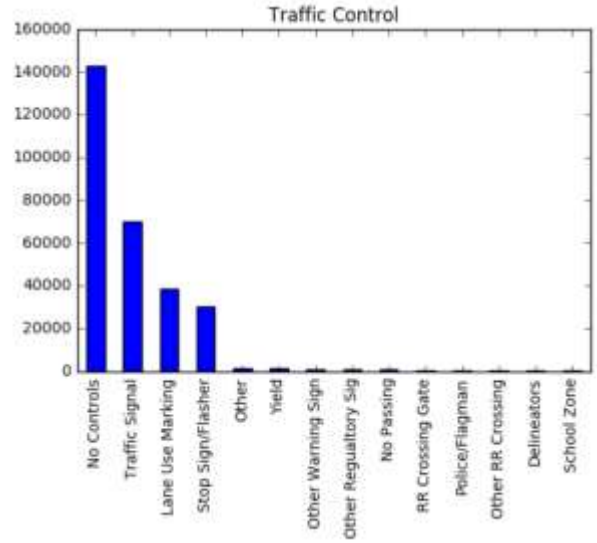

WEATHER
250000
200000
150000
100000
50000
0
Clear
Rain
Snow
Cloudy/Overcast
Fog/Smoke/Haze
Other
Sleet/Hail
Severe Cross Wind

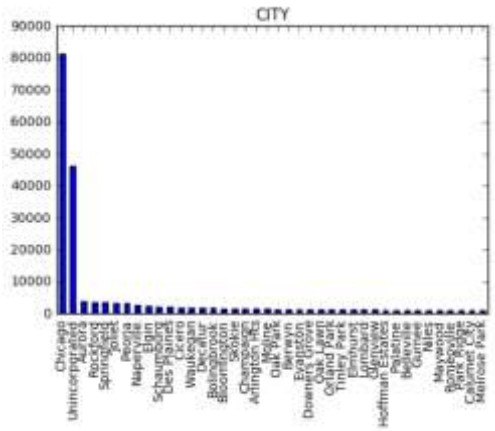

LIGHTING
200000
150000
100000
50000
0
Daylight
Darkness, Lighted Road
Darkness
Dusk
Dawn


Surface Condition
250000
200000
150000
100000
50000
0
Dry
Wet
Snow or Slush
Ice
Other
Sand, Mud, Dirt


Road Defect
250000
200000
150000
100000
50000
0
No Defects
Other
Rut, Holes
Debris On Roadway
Worn Surface
Shoulders


Road Feature
300000
250000
200000
150000
100000
50000
0
Not Applicable
Underpass
Bridge
(UNK)


Traffic Control
160000
140000
120000
100000
80000
60000
40000
20000
0
No Controls
Traffic Signal
Lane Use Marking
Stop Sign/Flasher
Other
Yield
Other Warning Sign
Other Regualtory Sig
No Passing
RR Crossing Gate
Police/Flagman
Other RR Crossing
Delineators
School Zone


CITY
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
Chicago

Evaluation

Performance Measure

The performance measure for this I chose is accuracy and AUC because this model predicts whether major accident would occur or not and for this reason evaluating on the basis of its accuracy is most important.

Classifiers

I used a number of Classifier but finally chose Gradiant Boosting Classifier. The parameters I fined tuned on were

- Number of trees : 500 as change after that is little and Smaller number of trees means Simpler Model
- Loss : deviance gives maximum accuracy and auc
- Learning Rate : .3 gives Peak AUC with almost best Accuracy
- Max Depth : 3 Both Accuracy and AUC reach maximum
- Sample Split : 2 Both Accuracy and AUC reach maximum
- Min Sample Leafs : 10 Both Accuracy and AUC reach maximum
- Sub Sample : 1 Both Accuracy and AUC reach maximum
- Max Features : auto

Evaluation Strategy

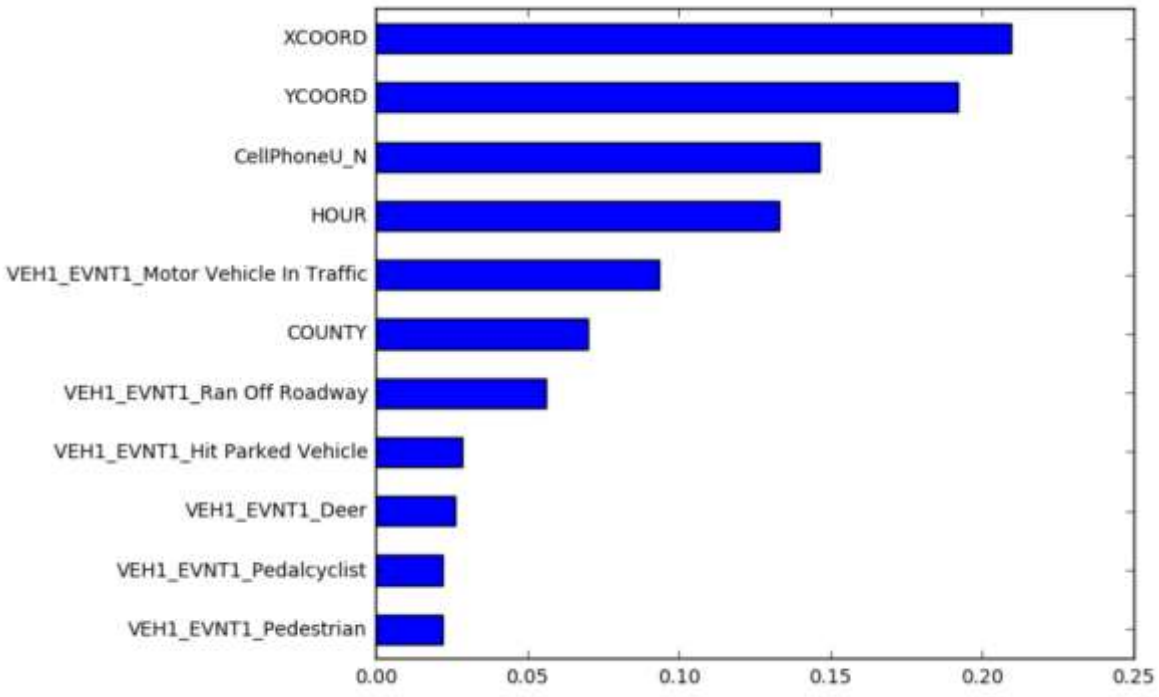
I used train-validation-test as I had a huge dataset and running cross validation on it would take a lot of time.

Performance Results

Model	**Parameters**	**Performance(Accuracy)**	Performance(AUC)
Baseline	majority class	78%	
Logistic regression		78.28%	64.14
Random Forest Classifier		84.75	76.45
Ada Classifier		87.47	88.094
Bagging Classifier		87.34	86.73
Extra Trees Classifier		86.5	84.79
Gradient Boosting		87.73	88.45
Gradient Boosting	Number of trees:500 Loss:deviance Learning Rate:.3 Max Depth: 3 Sample Split:2	88.4	89.41

	Min Sample Leafs:10 Sub Sample:1 Max Features:auto		

Top Features



Discussion

The results wrt to a random guess look good. The classifier I tried to optimized did not show much performance improvement. Maybe I should have tried optimizing other classfiers which were close to it without any optimization.

Interesting/Unexpected Results

The interesting observation from this model was that major accidents are located over particular region. This was derived as the top important features were coordinates. This could help analyze an mark certain zone for careful driving and limits.

Contributions of Each Group Member

I did it alone

Conclusion

The overall project was interesting and was fun to apply the learning got from class also some of my learning from Data Science knowledge before. This also gave an amazing insight on how major accidents and their cause are co related. What causes these what hour of the day and so on.