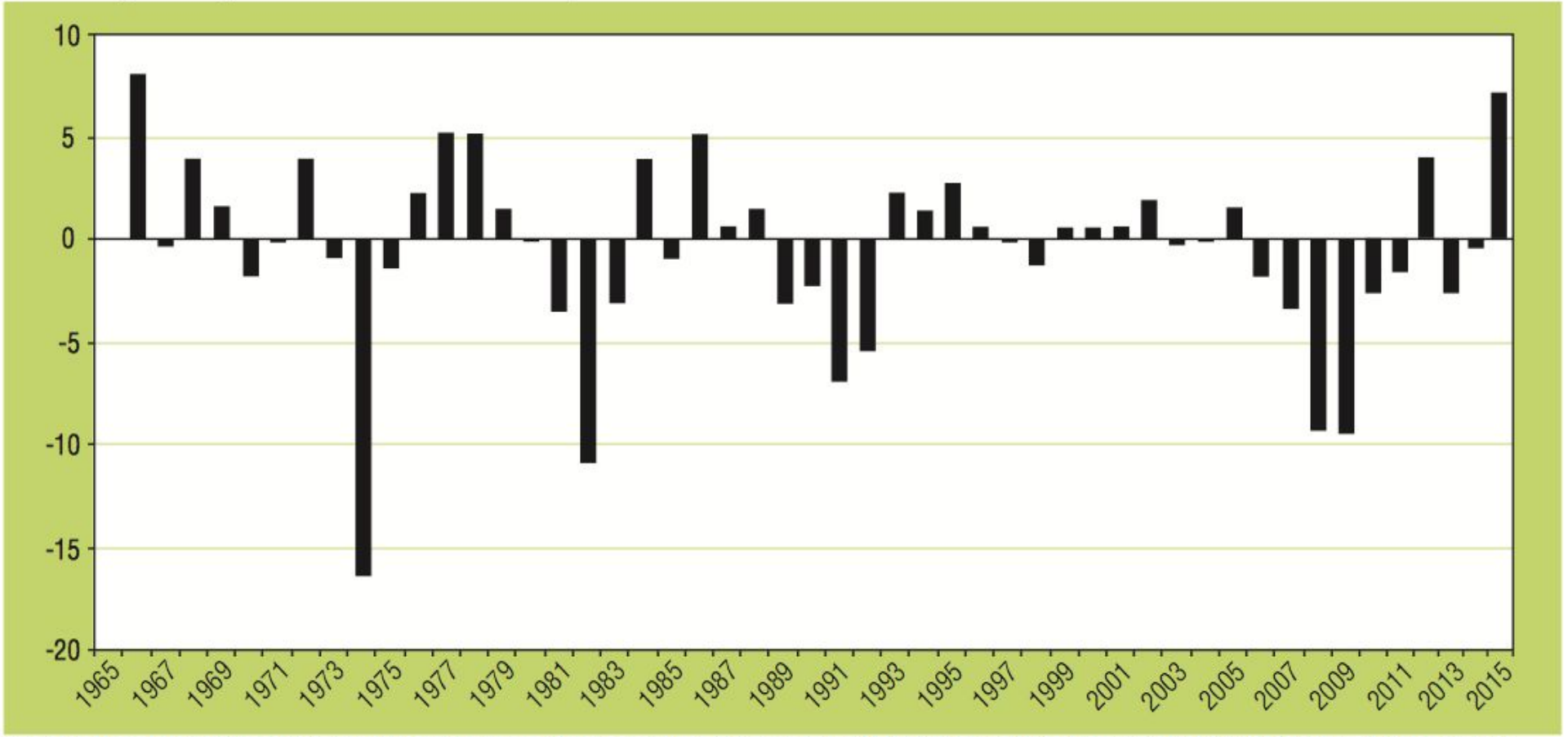
Potential Factors that Threaten Vehicular Safety

Sid Mamidanna
General Assembly
Data Science Immersive 2017

About Me

- Traffic management is among my three most enjoyed courses in college.
  - Traffic is fascinating. It’s a process with sprinkled unpredictability that we’re dependent on and need to be safe.
  - The opportunity to provide a useful insight on traffic with my abilities in data science, made me immediately decide to take on the traffic dataset.

Percentage Changes in accidents between 2004-2015.



Sources: 1965–1974: National Center for Health Statistics, HEW, and State Accident Summaries (Adjusted to 30-Day Traffic Deaths by NHTSA); FARS 1975-2014 Final File, 2015 Annual Report File (ARF)

Problem Statement

“Predicting crash survival of non negligent motor vehicle occupants.”

Definitions:

- Non Negligence: Aware that the driver is driving knowing they’re emotionally impaired, intoxicated, or driving a car with issues that could make them a liability in traffic.
- Motor Vehicle Occupants: Driver or passenger in a vehicle.

The Source of Data Set

This dataset comes from the NHTSA’s initiative to gather information on fatal crashes (FARS).

Fatal crashes are qualified as a crash involving a motor vehicle traveling on a public trafficway, which resulted in the death of a vehicle or a non-occupant within 30 days (720 hours) of the crash.

Table sources: law enforcement, public health, and transportation, among many other government entities.

Compiling a dataset to solve the problem

Accident: Crash characteristics and environmental conditions.
Vehicle: In-transport motor vehicle and occupant description.
Person: Information on the victims.

Table conditions: Exclude records or parked vehicles and reported drinking.

Limitations of the compiled data

Could use two other tables for a deeper understanding.

Removing null rows means a smaller training set.

Potentially missing correlated columns.

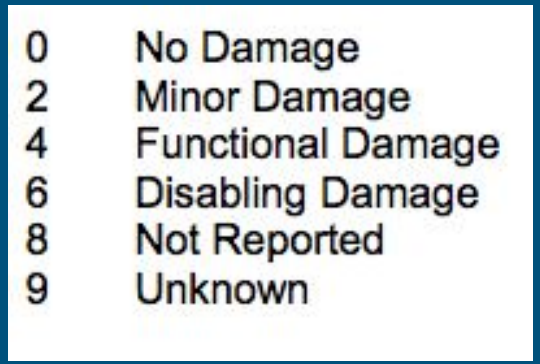
Original Table Size

Table	**Size (MB)**
Accident	5.4
Vehicle	13.3
Person	13.8
Total	150.5

Exploratory Data Analysis

- Metadata

Car Deformations

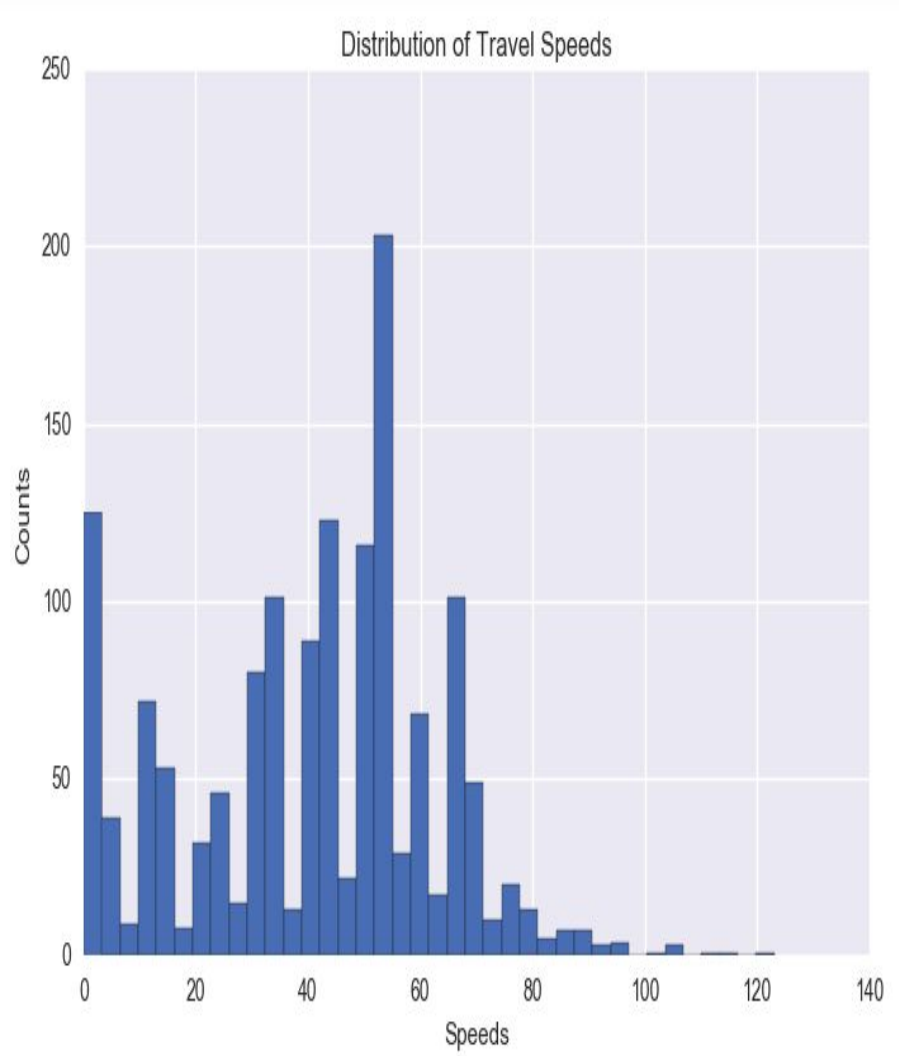


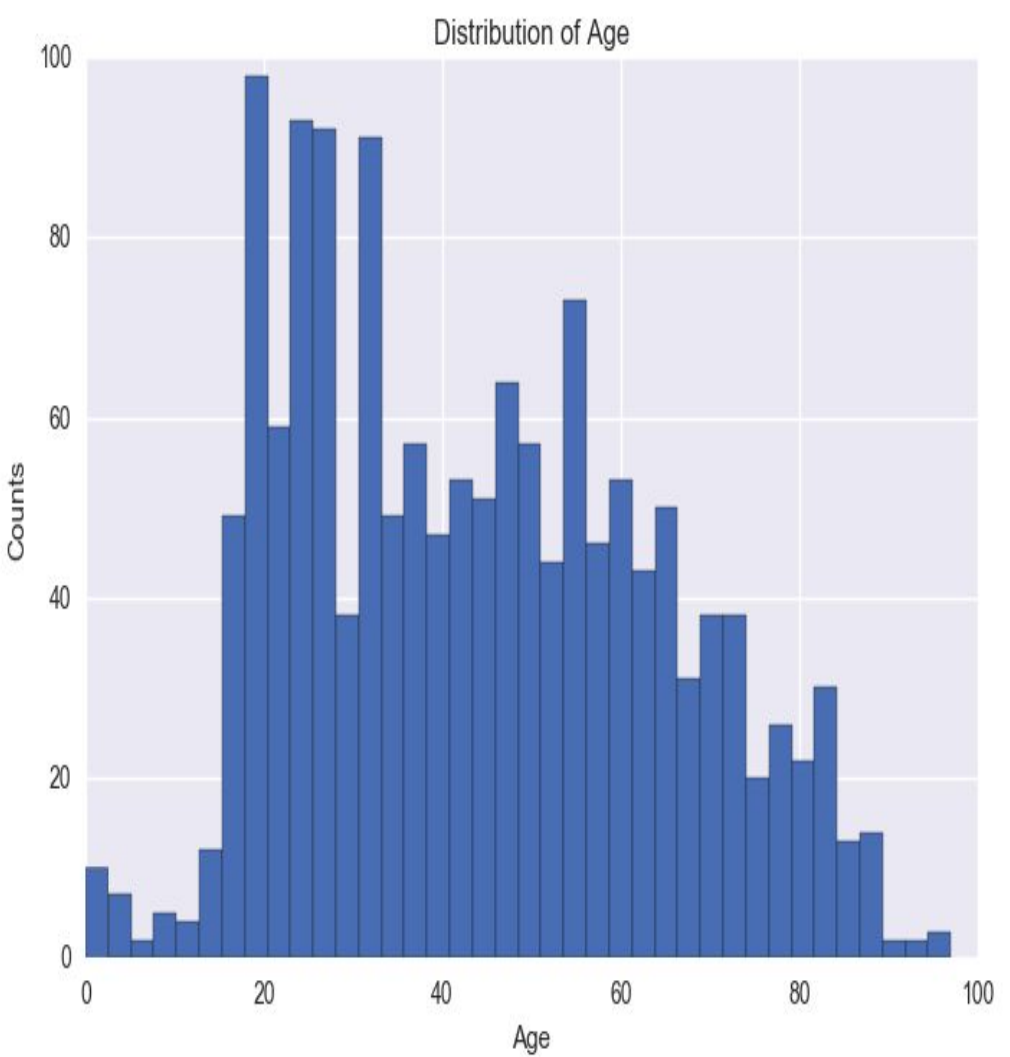
0	No Damage
2	Minor Damage
4	Functional Damage
6	Disabling Damage
8	Not Reported
9	Unknown

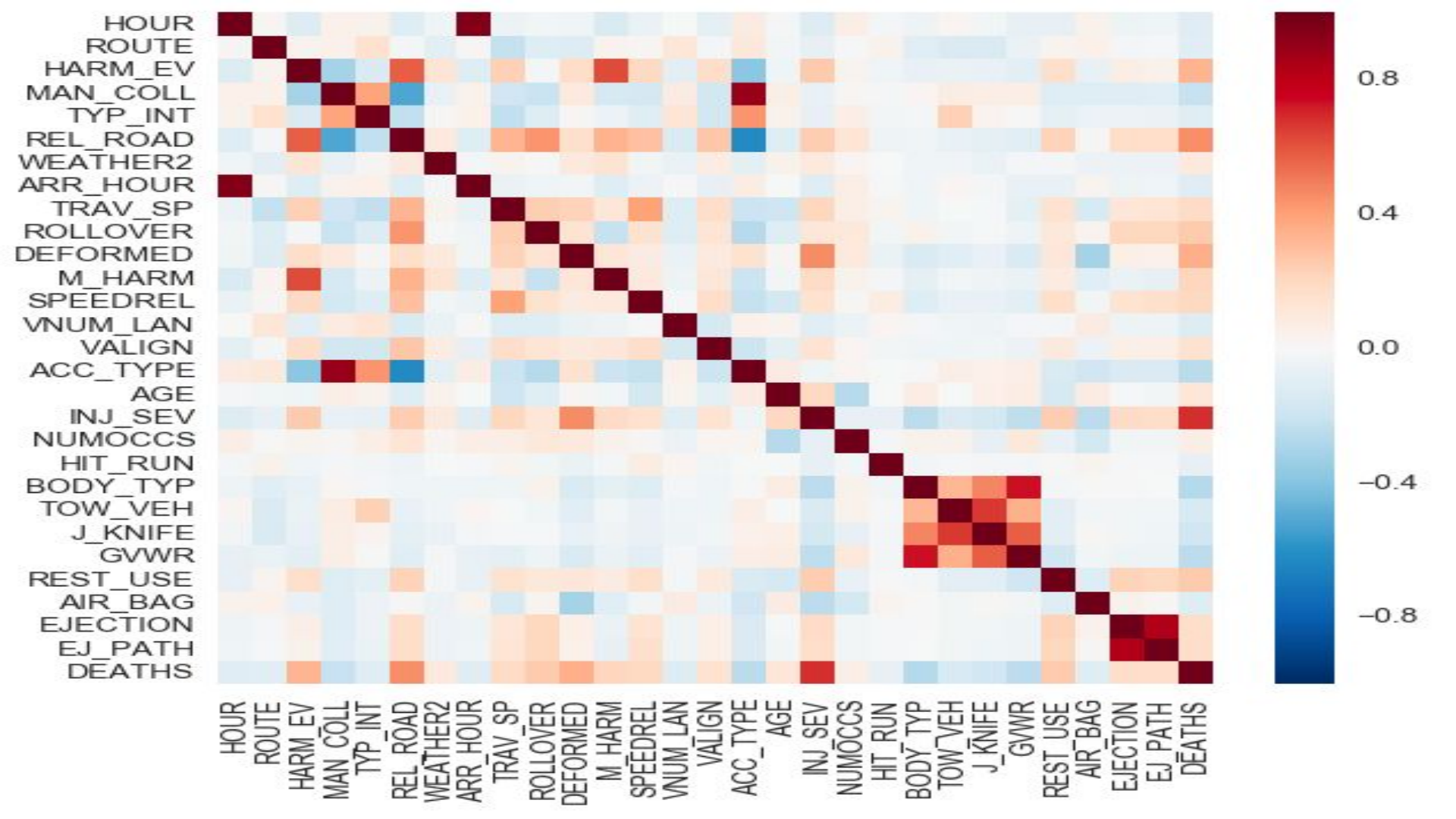
States

01	Alabama
02	Alaska
04	Arizona
05	Arkansas
06	California
08	Colorado
09	Connecticut
10	Delaware
11	District of Columbia
12	Florida
13	Georgia
15	Hawaii

- Arbitrary feature selection: determining which of the 65 variables to use.
- MVP: Most valuable predictors: HARM_EV, REL-ROAD, DEFORMED, INJ_SEV, BODY_TYP, REST_USE


Distribution of Travel Speeds
Counts
Speeds
0
50
100
150
200
250
0
20
40
60
80
100
120
140


Distribution of Age
Counts
Age
0
20
40
60
80
100
0
20
40
60
80
100


HOUR
ROUTE
HARM_EV
MAN_COLL
TYP_INT
REL_ROAD
WEATHER2
ARR_HOUR
TRAV_SP
ROLLOVER
DEFORMED
M_HARM
SPEEDREL
VNUM_LAN
VALIGN
ACC_TYPE
AGE
INJ_SEV
NUMOCCS
HIT_RUN
BODY_TYP
TOW_VEH
J_KNIFE
GVWR
REST_USE
AIR_BAG
EJECTION
EJ_PATH
DEATHS
0.8
0.4
0.0
−0.4
−0.8

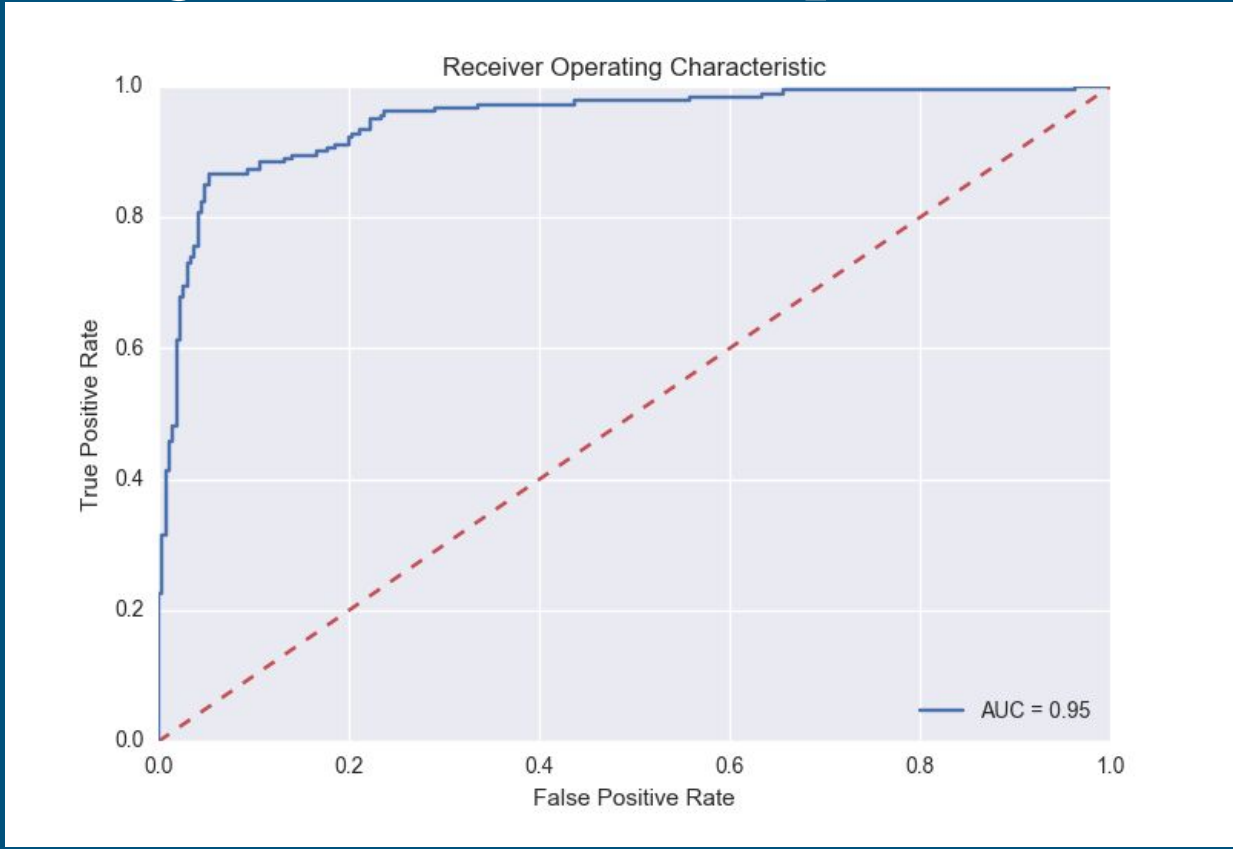
Models

- Logistic Regression
  - Plain, regularization techniques
  - Choice criteria: Quick computation, easy to interpret, and linear model assumption.
- Decision Tree
  - Plain and optimized
  - Choice Criteria: rulemaking
- SVM
  - Plain and optimized.
  - Choice criteria: Hopes of establishing a decision boundary.
- KNN
  - Plain and optimized.
  - Choice criteria: Hopes of establishing a decision boundary.

Model Comparison

Compare classification reports from each model. Describe reasons there were inconsistencies and how I marginalized it.
ROC CUrves

Logistic regression with 28 predictors

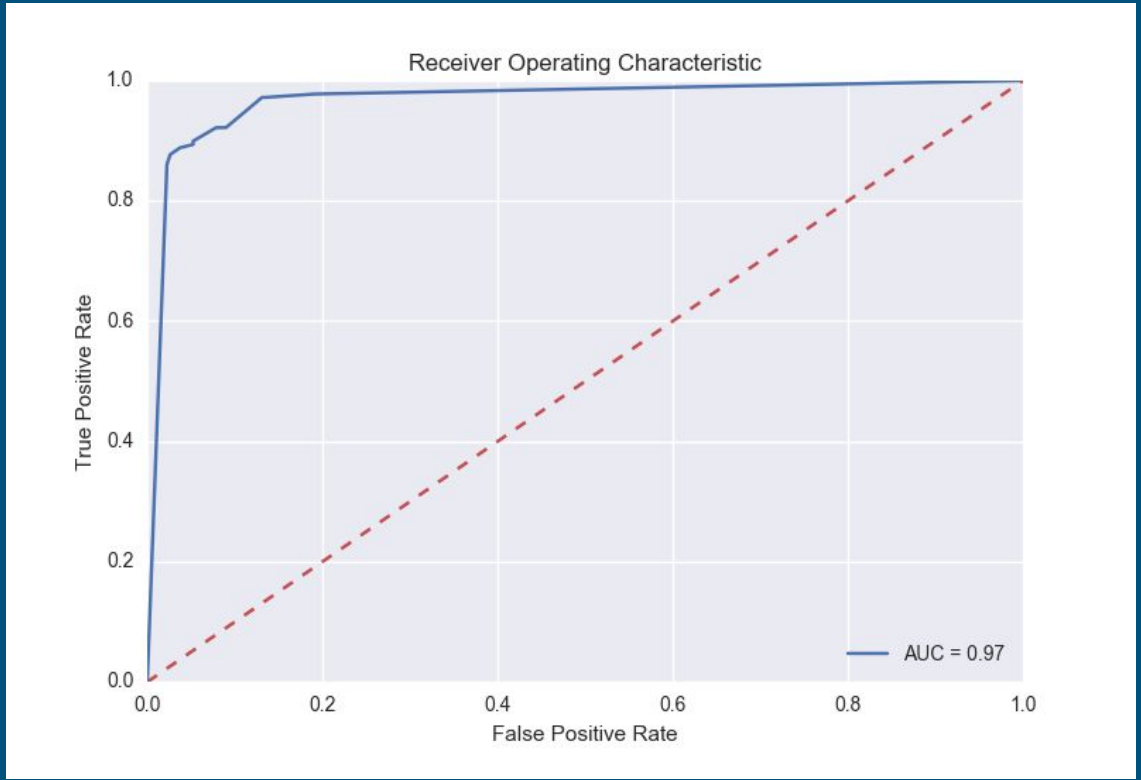


```
             precision    recall  f1-score   support

          0       0.92      0.89      0.90       280
          1       0.82      0.87      0.84       166

avg / total       0.88      0.88      0.88       446

[[248  32]
 [ 22 144]]
0.878923766816
```

```
t_logreg_all_vars = pd.DataFrame(sk.metrics.confusion_m
t_logreg_all_vars
```

	Predicted_0	Predicted_1
True_0	248	32
True_1	22	144

Decision tree with all 28 predictors

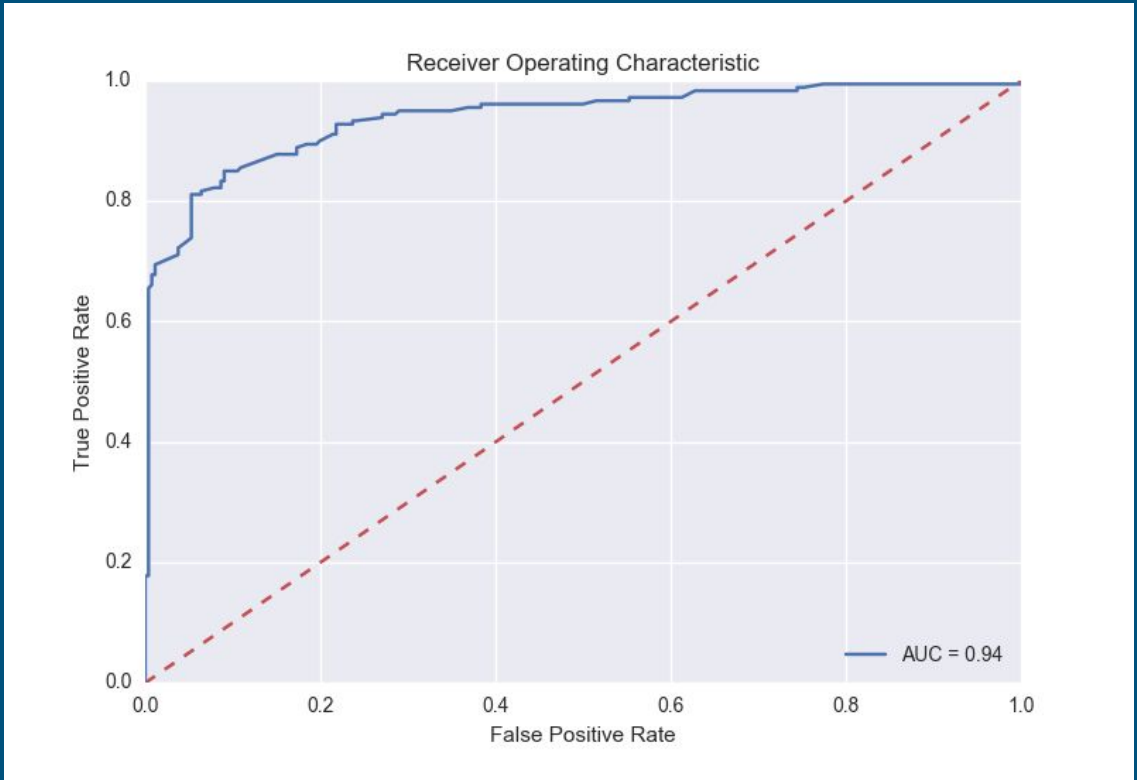


```
             precision    recall  f1-score   support

          0       0.69      0.99      0.81       273
          1       0.94      0.29      0.44       173

avg / total       0.79      0.72      0.67       446

[[270   3]
 [123  50]]
0.717488789238
```

```
cm_logreg_all_vars = pd.DataFrame(sk.metrics.confusi
cm_logreg_all_vars
```

	Predicted_0	Predicted_1
True_0	163	110
True_1	107	66

Logistic regression with five predictors

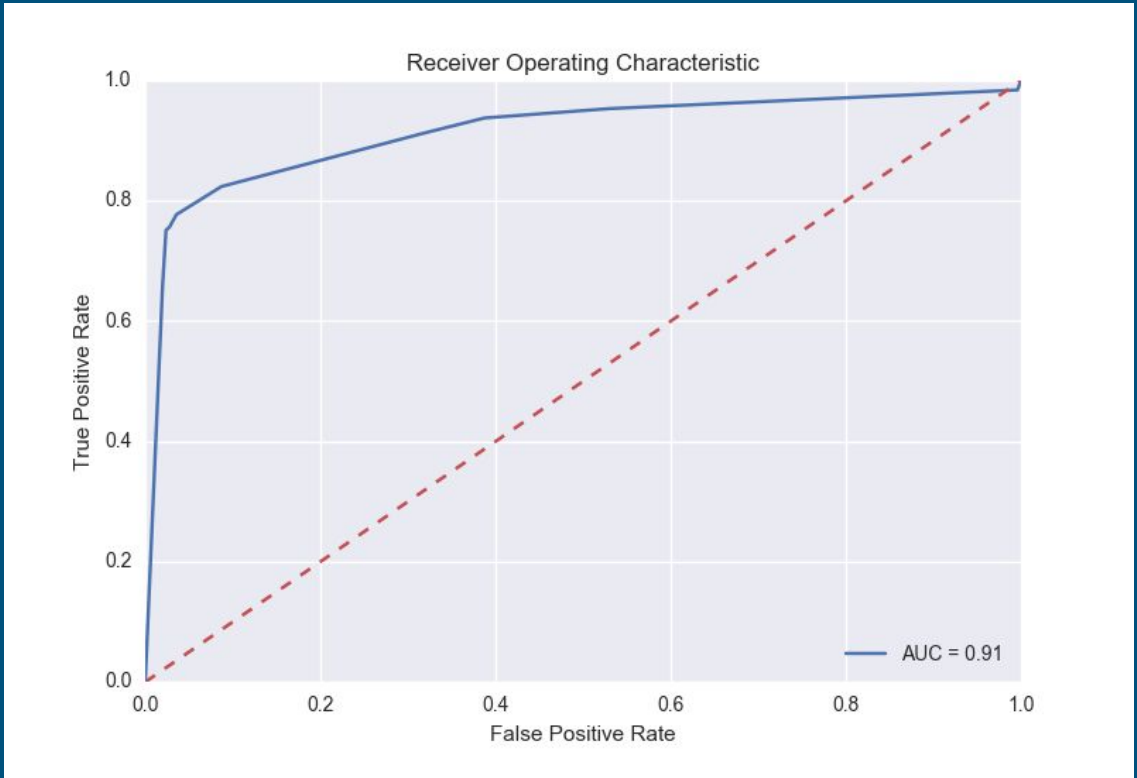


```
                precision    recall  f1-score   support

             0       0.93      0.89      0.91       283
             1       0.82      0.89      0.86       163

avg / total       0.89      0.89      0.89       446

[[252  31]
 [ 18 145]]
0.890134529148
```

```
cm_logreg_all_vars = pd.DataFrame(sk.metrics.confusion
cm_logreg_all_vars
```

	Predicted_0	Predicted_1
True_0	248	32
True_1	22	144

Decision tree with five predictors

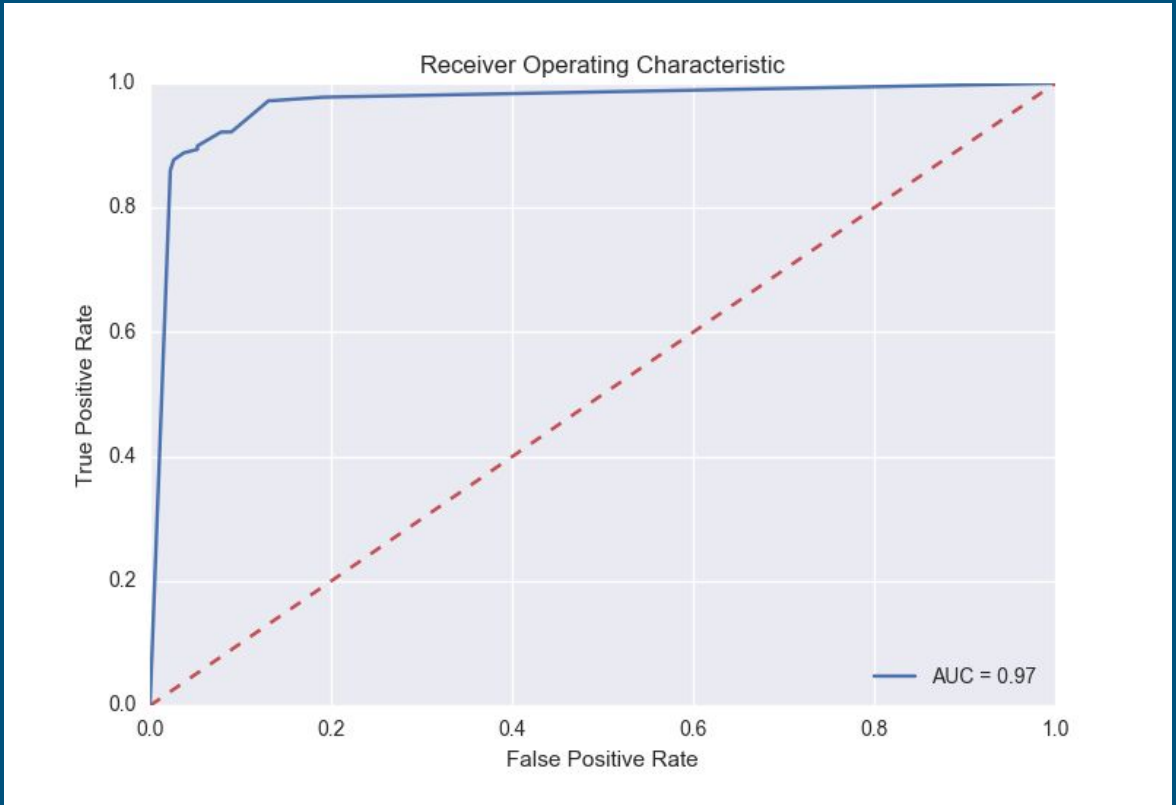


```
                precision    recall  f1-score   support

             0       0.85      0.99      0.92       258
             1       0.99      0.77      0.86       188

avg / total          0.91      0.90      0.89       446

[[256   2]
 [ 44 144]]
0.896860986547
```

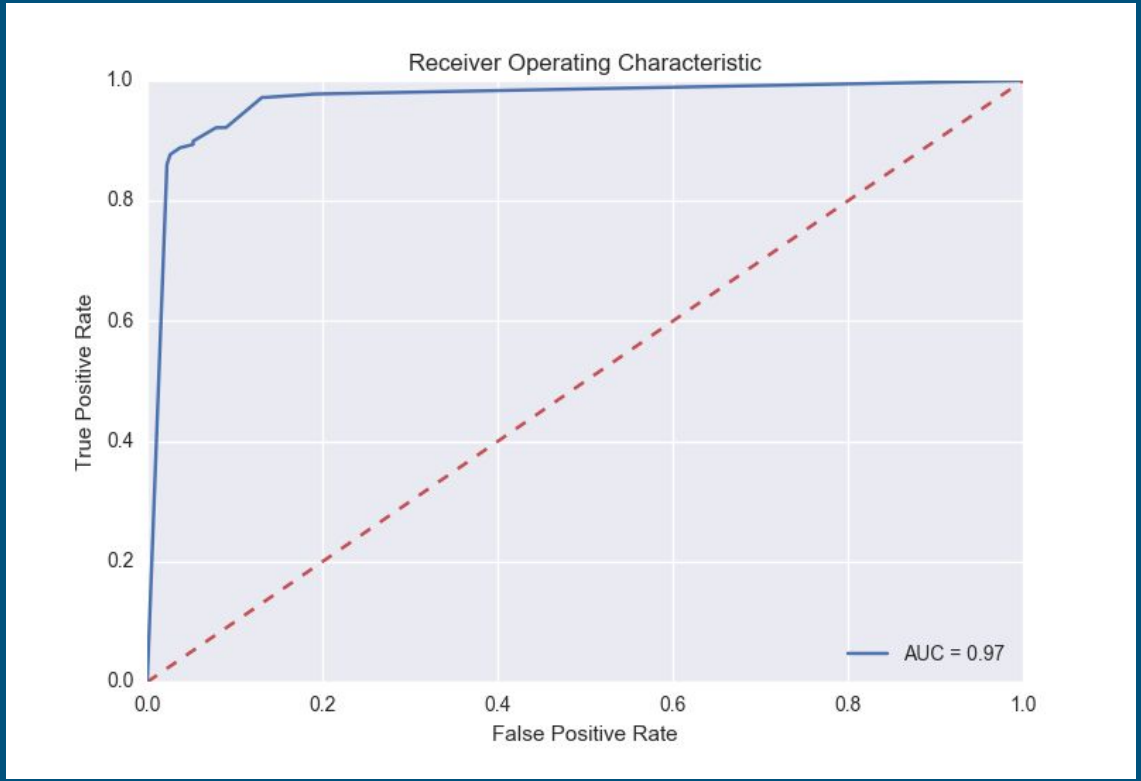
```
cm_logreg_all_vars = pd.DataFrame(sk.metrics.confusion
cm_logreg_all_vars
```

	Predicted_0	Predicted_1
True_0	163	95
True_1	107	81

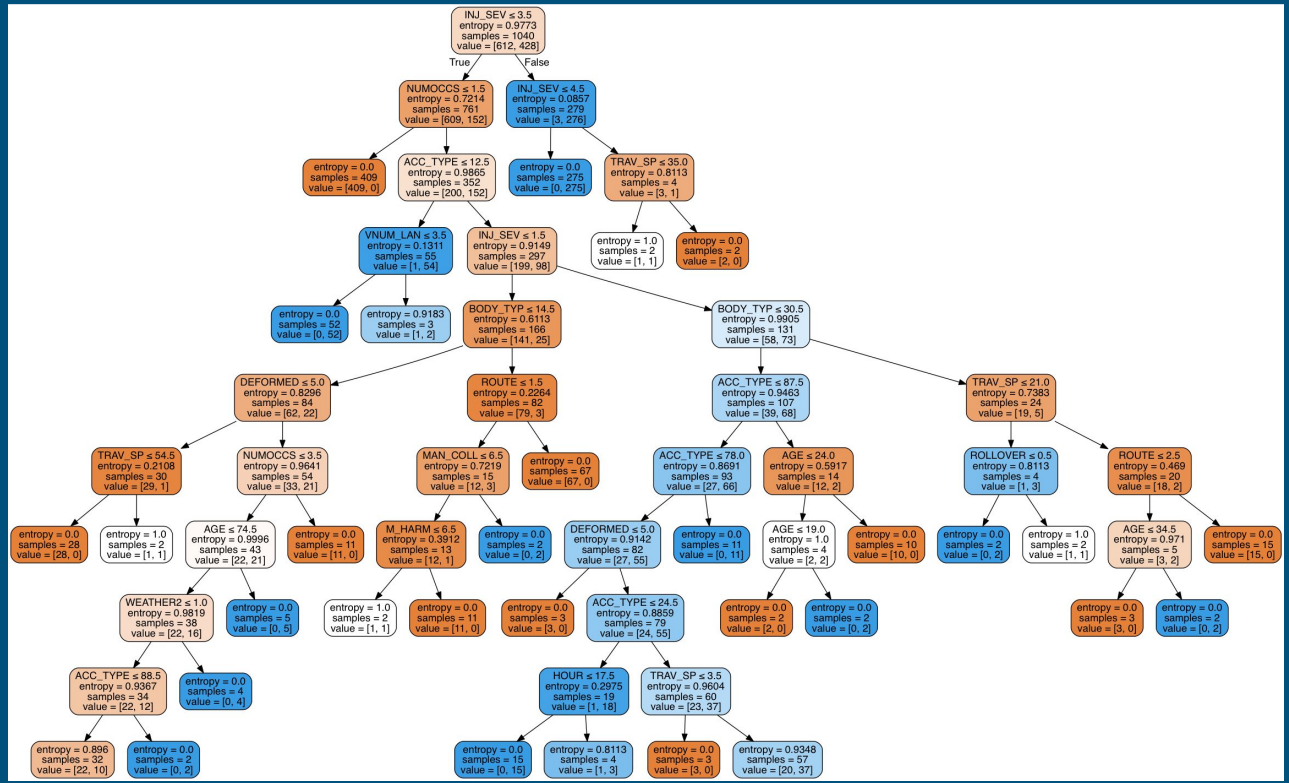
ROC Curve KNN

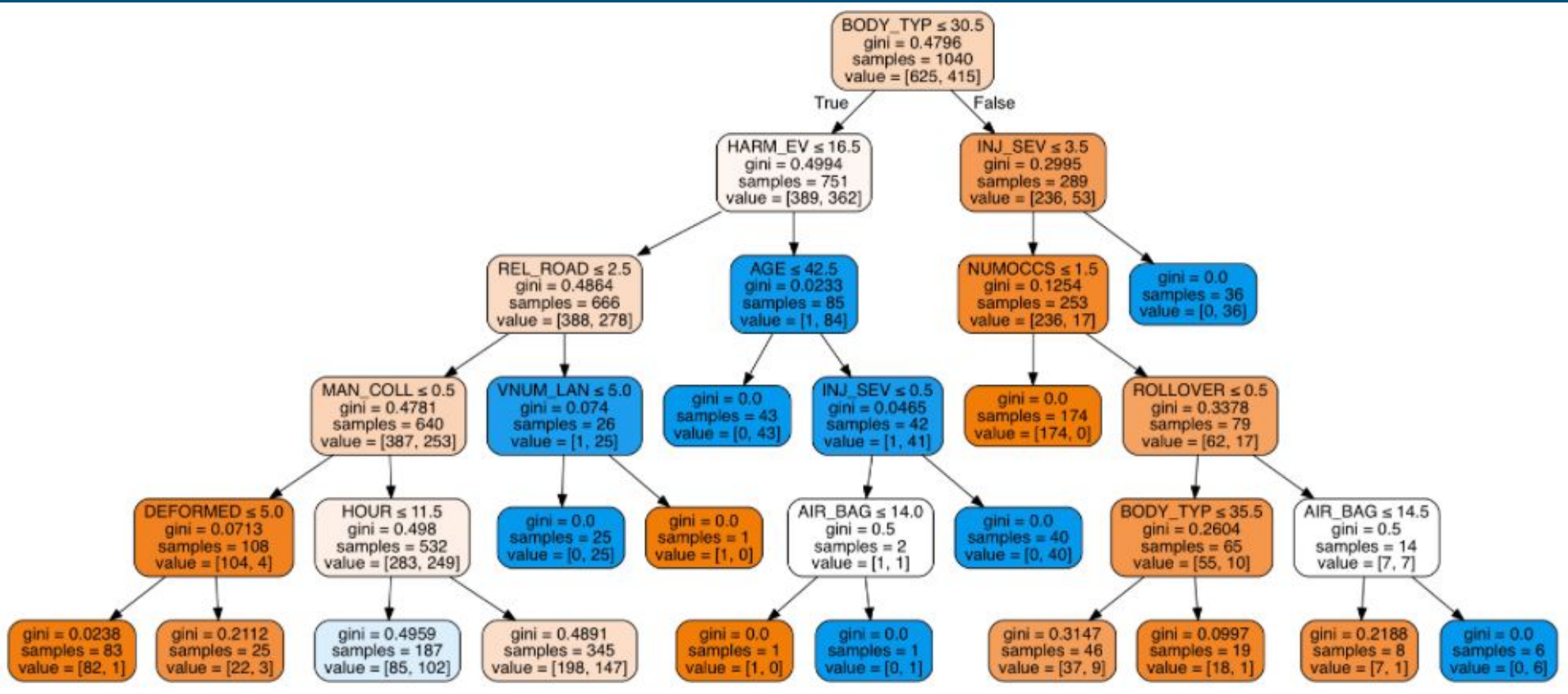


ROC Curve SVM

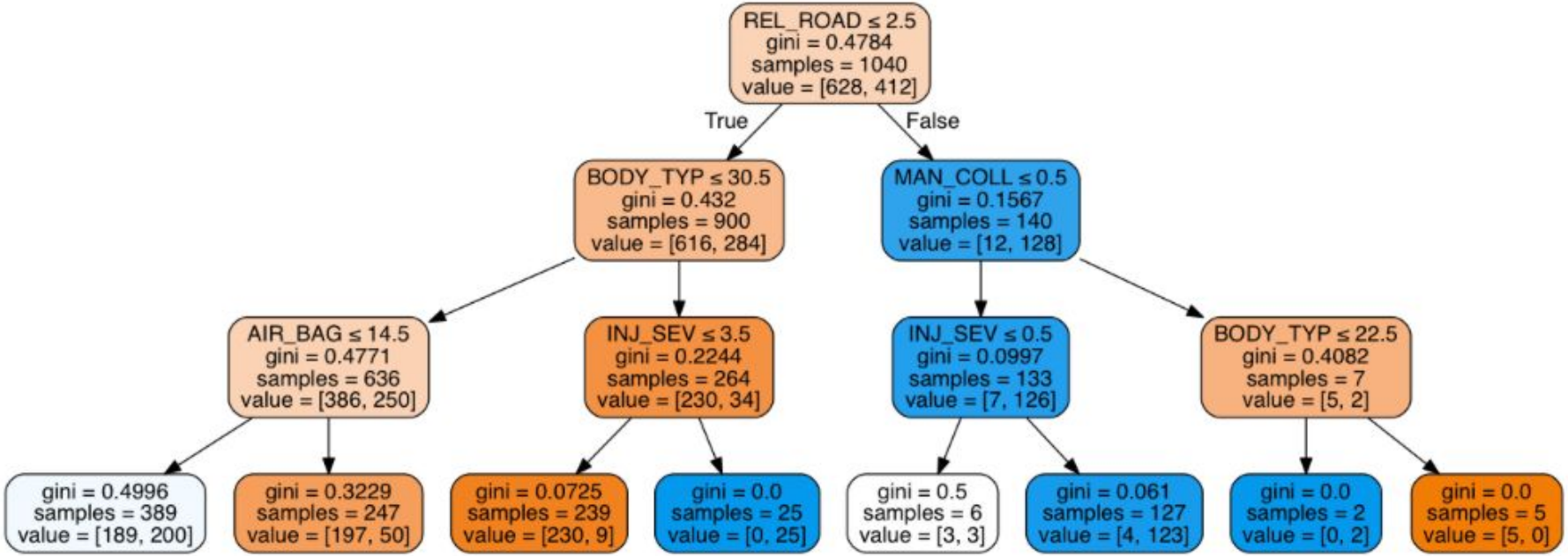


Decision Tree to go with




BODY_TYP ≤ 30.5
gini = 0.4796
samples = 1040
value = [625, 415]
True
False
HARM_EV ≤ 16.5
gini = 0.4994
samples = 751
value = [389, 362]
INJ_SEV ≤ 3.5
gini = 0.2995
samples = 289
value = [236, 53]
REL_ROAD ≤ 2.5
gini = 0.4864
samples = 666
value = [388, 278]
AGE ≤ 42.5
gini = 0.0233
samples = 85
value = [1, 84]
NUMOCCS ≤ 1.5
gini = 0.1254
samples = 253
value = [236, 17]
gini = 0.0
samples = 36
value = [0, 36]
MAN_COLL ≤ 0.5
gini = 0.4781
samples = 640
value = [387, 253]
VNUM_LAN ≤ 5.0
gini = 0.074
samples = 26
value = [1, 25]
gini = 0.0
samples = 43
value = [0, 43]
INJ_SEV ≤ 0.5
gini = 0.0465
samples = 42
value = [1, 41]
gini = 0.0
samples = 174
value = [174, 0]
ROLLOVER ≤ 0.5
gini = 0.3378
samples = 79
value = [62, 17]
DEFORMED ≤ 5.0
gini = 0.0713
samples = 108
value = [104, 4]
HOUR ≤ 11.5
gini = 0.498
samples = 532
value = [283, 249]
gini = 0.0
samples = 25
value = [0, 25]
gini = 0.0
samples = 1
value = [1, 0]
AIR_BAG ≤ 14.0
gini = 0.5
samples = 2
value = [1, 1]
gini = 0.0
samples = 40
value = [0, 40]
BODY_TYP ≤ 35.5
gini = 0.2604
samples = 65
value = [55, 10]
AIR_BAG ≤ 14.5
gini = 0.5
samples = 14
value = [7, 7]
gini = 0.0238
samples = 83
value = [82, 1]
gini = 0.2112
samples = 25
value = [22, 3]
gini = 0.4959
samples = 187
value = [85, 102]
gini = 0.4891
samples = 345
value = [198, 147]
gini = 0.0
samples = 1
value = [1, 0]
gini = 0.0
samples = 1
value = [0, 1]
gini = 0.3147
samples = 46
value = [37, 9]
gini = 0.0997
samples = 19
value = [18, 1]
gini = 0.2188
samples = 8
value = [7, 1]
gini = 0.0
samples = 6
value = [0, 6]

[1438]:


REL_ROAD ≤ 2.5
gini = 0.4784
samples = 1040
value = [628, 412]
True
False
BODY_TYP ≤ 30.5
gini = 0.432
samples = 900
value = [616, 284]
MAN_COLL ≤ 0.5
gini = 0.1567
samples = 140
value = [12, 128]
AIR_BAG ≤ 14.5
gini = 0.4771
samples = 636
value = [386, 250]
INJ_SEV ≤ 3.5
gini = 0.2244
samples = 264
value = [230, 34]
INJ_SEV ≤ 0.5
gini = 0.0997
samples = 133
value = [7, 126]
BODY_TYP ≤ 22.5
gini = 0.4082
samples = 7
value = [5, 2]
gini = 0.4996
samples = 389
value = [189, 200]
gini = 0.3229
samples = 247
value = [197, 50]
gini = 0.0725
samples = 239
value = [230, 9]
gini = 0.0
samples = 25
value = [0, 25]
gini = 0.5
samples = 6
value = [3, 3]
gini = 0.061
samples = 127
value = [4, 123]
gini = 0.0
samples = 2
value = [0, 2]
gini = 0.0
samples = 5
value = [5, 0]

Conclusions

- Though all four models performed similarly, it's best to go with the logistic model, since decision trees are sensitive to rules. Above, the decision tree diagrams look different simply because the maximum depth of branches was changed. Therefore, there's more backend work that's needed to provide a thorough suggestion with decision trees. Long and computationally extensive grid search is needed before providing any predictions based on decision trees.
- Discuss choice criteria under recommendation model.
- Future plans with this dataset - LOTS!

QUESTIONS