Capstone Project - Car Accident Severity

David Williams

Coursera Applied Data Science

**Introduction / Business Problem**

As more cars are on the road commuting to work from the suburbs, the traffic congestion is becoming a huge problem. Cities without mass public transportation are especially hit hard with bad traffic problems. As the number of vehicles increases, so also does traffic accidents. In addition to the number of vehicles, weather conditions, lighting, and road conditions play a significant part in traffic accidents.

A case study was conducted to analyze the conditions that cause automobile accidents. Using case study data, there needs to be a program developed to predict the severity of an accident based this historical data. Using the analysis, more emergency vehicles can be on standby, and drivers could be more careful when any of these conditions exist.

**Data**

The dataset contains 194,673 rows that contain historical data of prior accidents. There are 37 attributes in the dataset. The dependent variable is a column called SEVERITYCODE, which indicates the severity of the accident.

After loading the input file, the data was explored to determine the most appropriate columns to use as independent variables. Several Python methods were used to explore the data: the "head" method, the "info" method, and the "describe" method. The "unique" method was used on SEVERITYCODE to make sure it had binary values. The variable SEVERITYCODE was determined to be binary containing either the number 1 or 2 and verified that it is most appropriate for the dependent variable.

Several of the columns provide are in text format, which would make it difficult for machine learning. The columns that seem appropriate for machine learning are the numeric columns and the text columns that are codes or indicators. Some of the columns are inconsistent, having both numeric and charter data. The column UNDERINFL, which is, “whether or not a driver involved was under the influence of drugs or alcohol”, has some rows contain N, some rows contain 0, and some rows contain 1.

Having a binary dependent variable, a logistic regression model will be built using SEVERITYCODE as the dependent variable. From the exploration of the data, the six variables in table 1 were chosen as the independent variable for the model.

Table 1

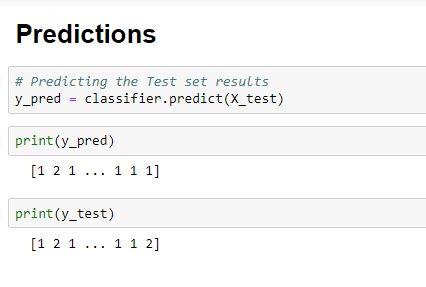
Dependent and Independent variables

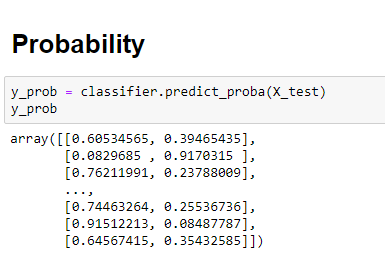
|  |  |
| --- | --- |
| SEVERITYCODE | A code that corresponds to the severity of the collision: 1 or 2 |
| PERSONCOUNT | The total number of people involved in the collision |
| PEDCOUNT | The number of pedestrians involved in the collision. This is entered by the state |
| PEDCYLCOUNT | The number of bicycles involved in the collision. This is entered by the state |
| VEHCOUNT | The number of vehicles involved in the collision. This is entered by the state |
| SDOT\_COLCODE | A code is given to the collision by SDOT  Appendix B describe the codes |
| HITPARKEDCAR | Whether or not the collision involved hitting a parked car (Y/N) |

**Methodology**

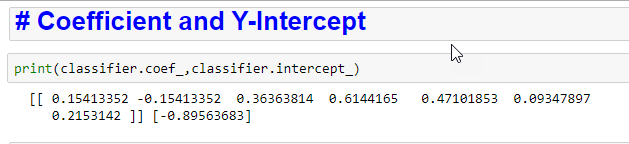
The steps for developing the model

1. The dependent and independent variables were copied into a separate data frame.
2. The rows that had columns with missing data were dropped.
3. The independent variables were copied into the X matrix and the dependent variable copied in the Y matrix.
4. Created visualization of the independent numeric variables. Visualization is in the appendix A.
5. The variable HITPARKCAR is a character field (Y/N). It was transform using LabelEncoder and OneHotEncoder.
6. The X matrix and Y matrix was split into test and train data.
7. Feature Scaling was run on the independent variables.
8. The logistic classifier was created.
9. Prediction and probability were ran using the X\_test. Y\_test was compared to the Y\_pred.



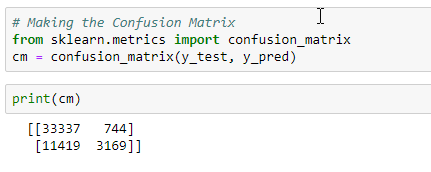


1. Coefficient and y-intercept were created.



**Discussion**

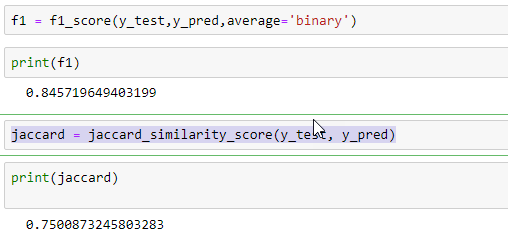
Confusion matrix show 33,337 true positive, 3,169 true negative, 11,419 false negative, and 744 false positive.



F1 is an overall measure of a model's accuracy that combines precision and recall. A good F1 score means that you have low false positives and low false negatives, so you're similarity correctly identifying real threats, and you are not disturbed by false alarms. An F1 score is considered perfect when it's 1, while the model is a total failure when it's 0.

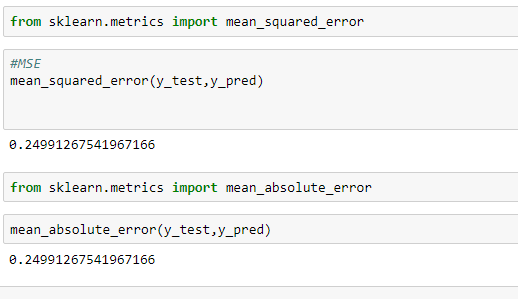
F1 = 2 \* (precision \* recall) / (precision + recall)

The Jaccard similarity index (sometimes called the Jaccard similarity coefficient) compares members for two sets to see which members are shared and which are distinct. It's a measure of for the two sets of data, with a range from 0% to 100%. The higher the percentage, the more similar the two populations.



The mean square error is the average of the square of the difference between the observed and predicted values of a variable.

The mean absolute error is a measure of errors between paired observations expressing the same phenomenon.



**Results**

The results in a logistic model can be written using the y-intercept and coefficient:

SEVERITYCODE = -0.89563683

+ 0.15413352 \* HITPARKEDCAR

+ 0.36363814 \* PERSONCOUNT

+ 0.6144165 \* PEDCOUNT

+ 0.47101853 \* PEDCYLCOUNT

+ 0.09347897 \* VEHCOUNT

+ 0.2153142 \* SDOT\_COLCODE

**Conclusion**

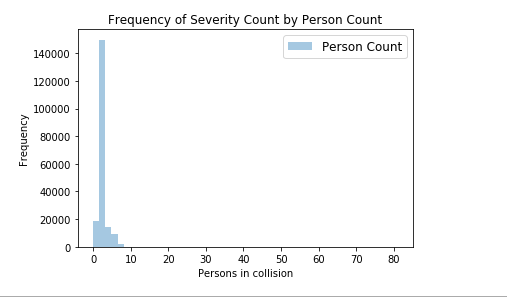
The coefficient of the linear equation indicates the magnitude of each of the variables. The y-intercept is the value if all other factors do not exist.

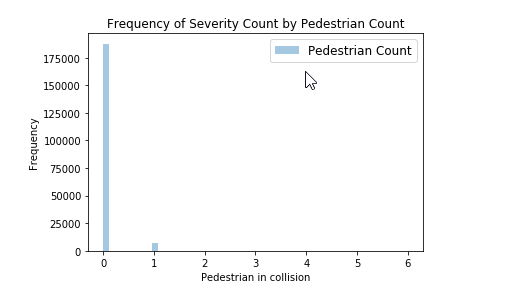
Based on the coefficient’s, pedestrian accidents are the most significant factor in the

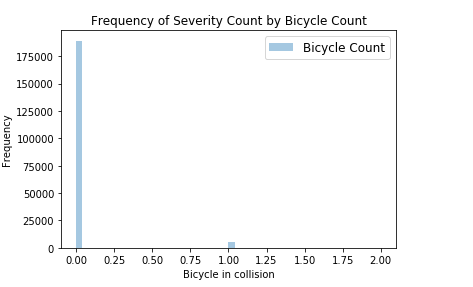
severity of the collision. Most collisions are two vehicle collisions and from the

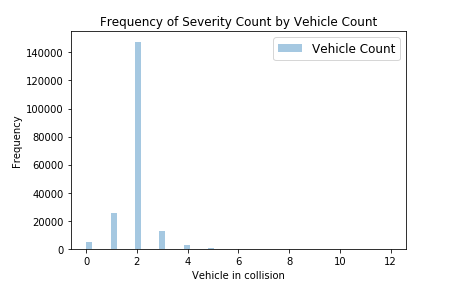
SDOT\_COLCODE (10), the most collisions are “Entering At Angle” .

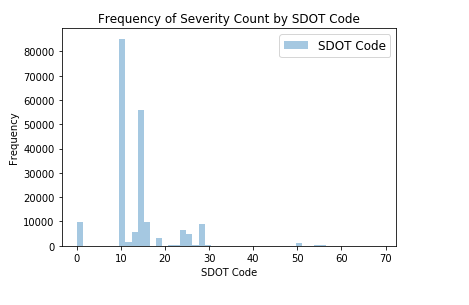
Appendix A











Appendix B

|  |  |
| --- | --- |
| **Code** | **Description** |
| 0 | Vehicle Going Straight Hits Pedestrian |
| 1 | Vehicle Turning Right Hits Pedestrian |
| 2 | Vehicle Turning Left Hits Pedestrian |
| 3 | Vehicle Backing Hits Pedestrian |
| 4 | Vehicle Hits Pedestrian - All Other Actions |
| 5 | Vehicle Hits Pedestrian - Actions Not Stated |
| 10 | Entering At Angle |
| 11 | From Same Direction -Both Going Straight -Both Moving- Sideswipe |
| 12 | From Same Direction -Both Going Straight - One Stopped- Sideswipe |
| 13 | From Same Direction - Both Going Straight - Both Moving - Rear End |
| 14 | From Same Direction - Both Going Straight - One Stopped - Rear End |
| 15 | From Same Direction - One Left Turn - One Straight |
| 16 | From Same Direction - One Right Turn - One Straight |
| 19 | One Car Entering Parked Position |
| 20 | One Car Leaving Parked Position |
| 21 | One Car Entering Driveway Access |
| 22 | One Car Leaving Driveway Access |
| 23 | From Same Direction - All Others |
| 24 | From Opposite Direction - Both Moving - Head On |
| 25 | From Opposite Direction - One Stopped - Head On |
| 26 | From Opposite Direction - Both Going Straight -sideswipe |
| 27 | From Opposite Direction - Both Going Straight - One Stopped - sideswipe |
| 28 | From Opposite Direction - One Left Turn - One Straight |
| 29 | From Opposite Direction - One Left Turn - One Right Turn |
| 30 | From Opposite Direction - All Others |
| 31 | Not Stated |
| 32 | One Parked - One Moving |
| 40 | Train Struck Moving Vehicle |
| 41 | Train Struck Stopped or Stalled Vehicle |
| 42 | Vehicle Struck Moving Train |
| 43 | Vehicle Struck Stopped Train |
| 44 | Unicycle |
| 45 | Bicycle |
| 46 | Tricycle |
| 47 | Domestic Animal (horse, cow, sheep, etc) |
| 48 | Domestic Animal Other (Cat, Dog etc) |
| 49 | Non Domestic Animal (deer, bear, elk, etc) |
| 50 | Struck Fixed Object |
| 51 | Struck Other Object |
| 52 | Vehicle Overturned |
| 53 | Person Fell, Jumped, or was Pushed From Vehicle |
| 54 | Fire Started In Vehicle |
| 55 | Accidently Overcame By Carbon Monoxide Poison |
| 56 | Breakage Of Any Part Of the Vehicle Resulting Injury or in Further Property Damage |
| 57 | All Other Non-Collisions |
| 60 | Vehicle Hits State Road or Construction Machinery |
| 61 | Vehicle Struck By State Road or Construction Machinery |
| 62 | Vehicle Hits County Road or Construction Machinery |
| 63 | Vehicle Struck By County Road or Construction Machinery |
| 64 | Vehicle Hits City Road or Construction Machinery |
| 65 | Vehicle Struck By City Road or Construction Machinery |
| 66 | Vehicle Hits Other Road or Construction Machinery |
| 67 | Vehicle Struck by Other Road or Construction Machinery |
| 71 | Same Direction - Both Turning Right - Both Moving - Sideswipe |
| 72 | Same Direction - Both Turning Right - One Stopped - Sideswipe |
| 73 | Same Direction - Both Turning Right - Both Moving -Rear End |
| 74 | Same Direction - Both Turning Right - One Stopped -Rear End |
| 81 | Same Direction - Both Turning Left - Both Moving -Sideswipe |
| 82 | Same Direction - Both Turning Left - One Stopped -Sideswipe |
| 83 | Same Direction - Both Turning Left - Both Moving -Rear End |
| 84 | Same Direction - Both Turning Left - One Stopped -Rear End |