PROJECT REPORT

GANESH P UMARANI

Under the guidance of:

PROF. SANJUKTA DAS SMITH

OBJECTIVE:

The primary objective of this project is to find, analyze the various causes of fatal accidents in USA from 2010 to 2017and formulate them towards law making or highway safety rules and regulations.

GOALS:

Analyzing high level overview that can be used to answer a multitude of questions concerning the safety of vehicles, drivers, traffic situations, roadways, and environmental conditions. Some specific policy and research can be achieved which include:

- Alcohol-related legislation,
- Motorcycle helmet legislation,
- Restraint usage legislation,
- Speed limit laws,
- Vehicle safety designs,
- Large-truck safety, and
- Air bag effectiveness.

METHODOLOGIES:

This problem statement ideally needs to follows the data science procedures and there is a lot of learning involved in each of these steps:

1. Data Acquisition
2. Data integration

3. Data cleaning
4. Data preprocessing
5. Dimensionality reduction
6. Modelling and Testing
7. Challenges

Data Acquisition:

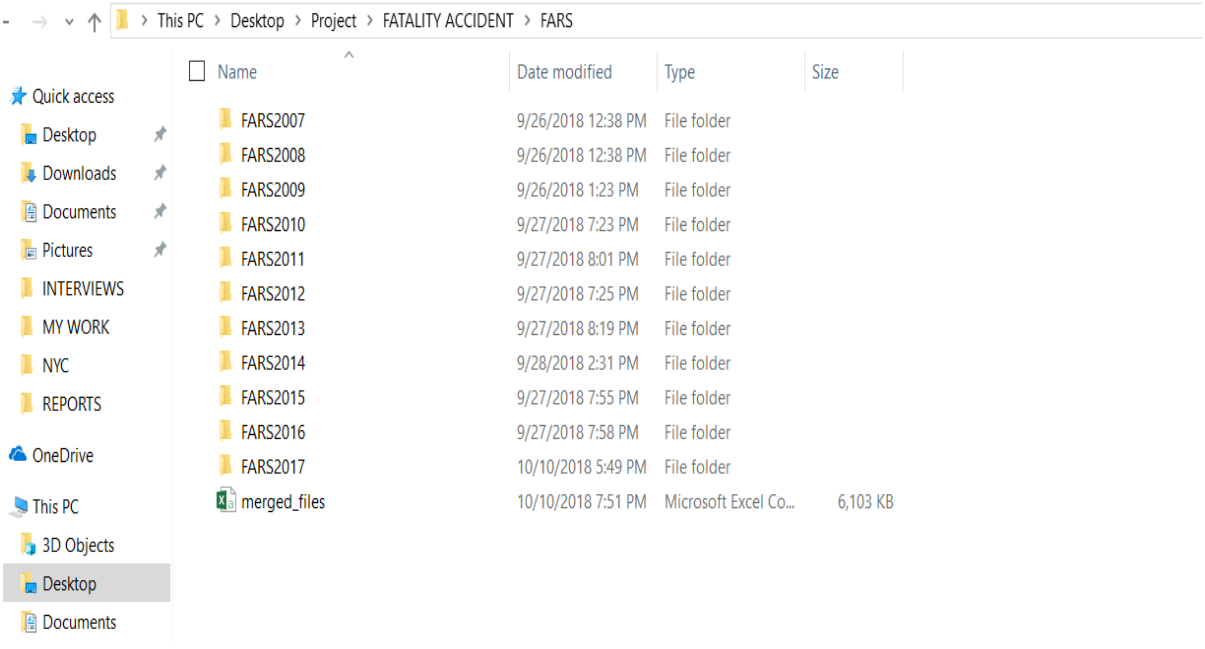
The data is available from National Highway Traffic Safety Administration's Fatality Analysis Reporting System (FARS) which is available here from year 1975 to 2017. But taking into consideration many underlying factors that might have changed across years like Highway rules and regulations, vehicles restrictions & safety rules and to reduce the broad spectrum of data, we have considered data from year 2010 to 2017 for analysis purpose.

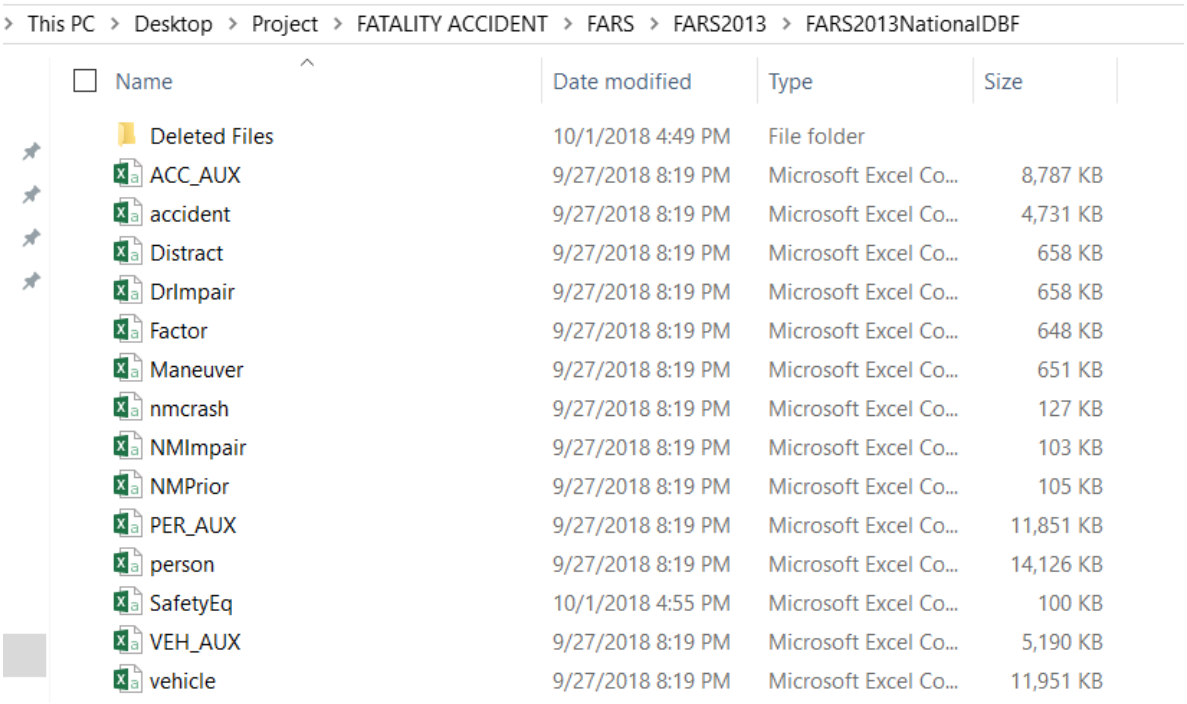
The Fatality Analysis Reporting System (FARS) contains data derived from a census of fatal traffic crashes within the 50 States, the District of Columbia, and Puerto Rico. The FARS database contains descriptions, in a standardized format, of each fatal crash reported. Data comes primarily from the police accident report (PAR) in that State, but also from death certificates, State coroners and medical examiners, State driver and vehicle registration records, and emergency medical services records.

Data Integration:

Each year has set of many files which describe few attributes like accident type, person, driver's information etc. First step involves selecting only those files which are contributing for the analysis. So all the files are critically analyzed and favorable ones are selected. The selected files are: Accident, Person, Vehicle, Distract, Drimpair, Factor, Maneuver, nmcrash, nmimpair, nmprior, VINDecode, Vision, and SafetyEq. These files are merged horizontally which gives around 385000 records and around 225 variables. This is handled such that no two files have redundant features. Also many features are discarded which are either irrelevant or duplicated which brought the variables down to 112. For each year, we bring down the number of files from around 15 to 3 parent files – ACCIDENT, PERSON and VEHICLE. These files across years are merged vertically handling many conflicts, variable name inconsistency and many redundant variables are

ignored and removed. This helped in getting the data in simple and uniform pattern. The primary target variable is DEATH so deaths involved in the accidents are available in ACCIDENT file. Hence from here forth the analysis turns completely on ACCIDENT file. The dataset to be used in further steps has 341685 records and 64 variables. This dataset is inspected further for blank, missing, NaN values and other inconsistency issues.





Data Cleaning:

Cleaning is done using both Python and Excel. This step involves various cleaning procedures:

- Remove variables with more than 75% of NaN or blank values:
  MAKE has more than 75% of NaN values so the variables was removed.
- Drop records with missing values:
  Examining missing values is important because when some of your data is missing, the data set can lose expressiveness, which can lead to weak or biased analyses. Practically, this means that when you're missing values for certain features, the chances of your classification or predictions for the data being off only increase.
  Some records has blank values which were removed. Data imputation was considered irrelevant in this case as mostly all the variables are categorical and data imputation does not perform well in that case.
- Drop records with null values :
  Similarly, null values were eliminated.
- Removing unknowns:
  Many of the variables have unknown values. Unknown is yet another category. It can represent either missing value or something which is different and does not belong to the group or cannot be defined.
- Renaming columns to a more recognizable set of labels

Data Preprocessing:

- The first step for preprocessing is to factorize the variables which are categorical in nature.
- Then the next step is to check for all the datatypes of the variables. The dataset contains 80% of categorical variables- nominal and ordinal, 15% are count variables and rest are numerical. Thus all the variables are assigned their specific datatype in python.
- The DEATH variable which is a count variable is converted for binary (the reason for this is explained in further steps).

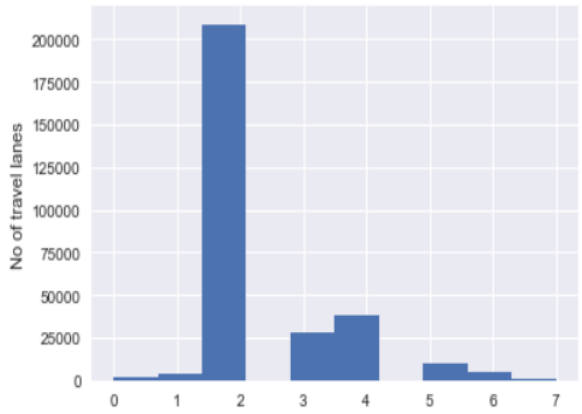
- Fixing up formats:
  Often the data is loaded or saved, some data might not be in correct formats. The variable LAST_YR was incorrectly formatted for few records. Thus the whole column is formatted to fix this issue. A typical job when it comes to cleaning data is correcting these types of issues.

UNIVARIATE ANALYSIS:

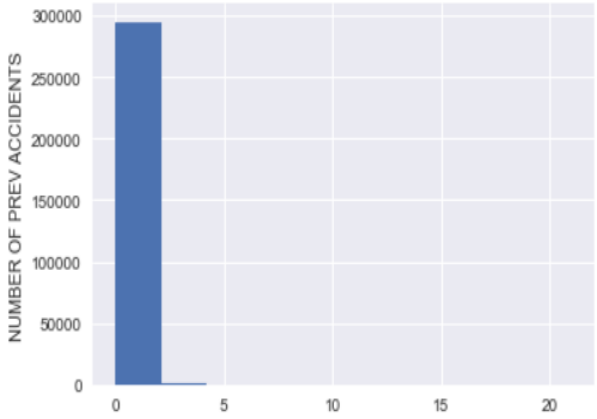
We can see that most of the count variables are following the Poisson distribution. Also the mean and variance in all the cases are approximately equal. Thus DEATHS can be approximated by a Poisson distribution.

The count variables in the dataset are as follows

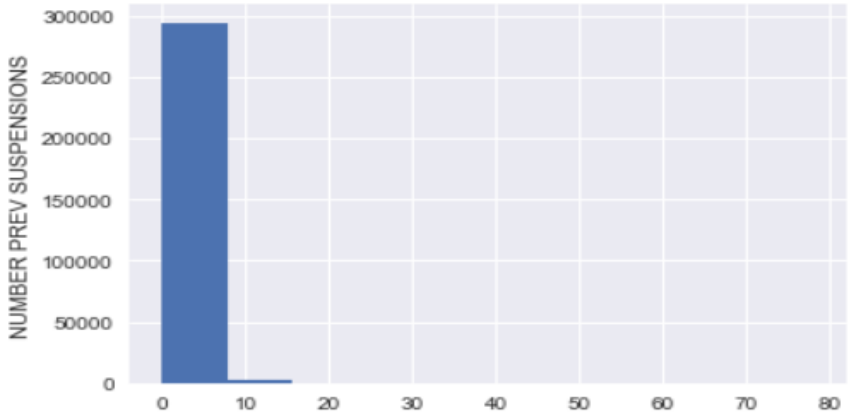
NUMBER OF TRAVEL LANES
[2 4 3 6 5 0 1 7]



NUMBER OF PREV ACCIDENTS
[0 1 2 3 4 5 18 6 7 14 21 8 10 9]

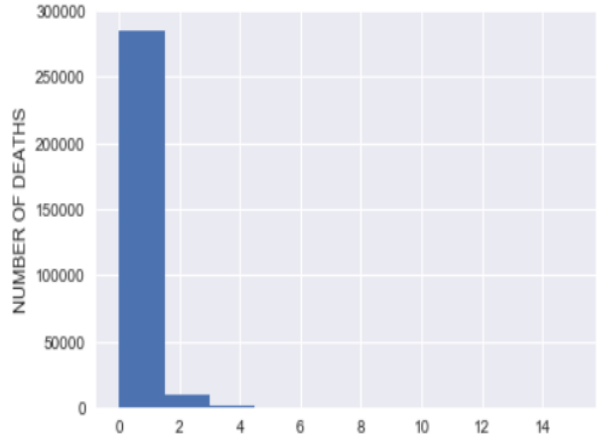


NUMBER PREV SUSPENSIONS
[0 1 4 2 3 6 12 5 9 8 7 14 15 10 16 13 11 24 20 27 18 17 55 19
44 29 21 26 22 39 25 28 23 31 30 33 65 42 38 34 52 59 35 32 78 69 60 46
47 75 63 45 36 51 37 43]



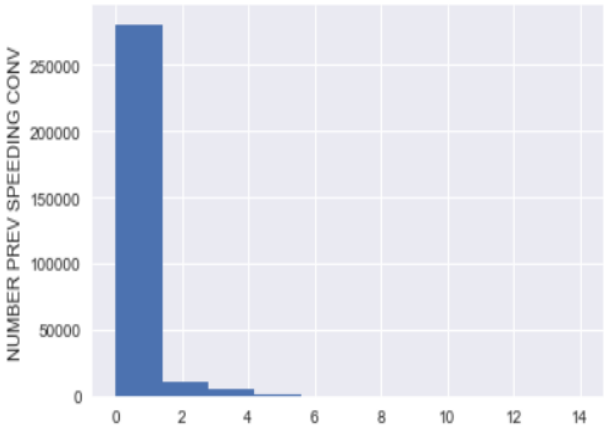
NUMBER OF DEATHS

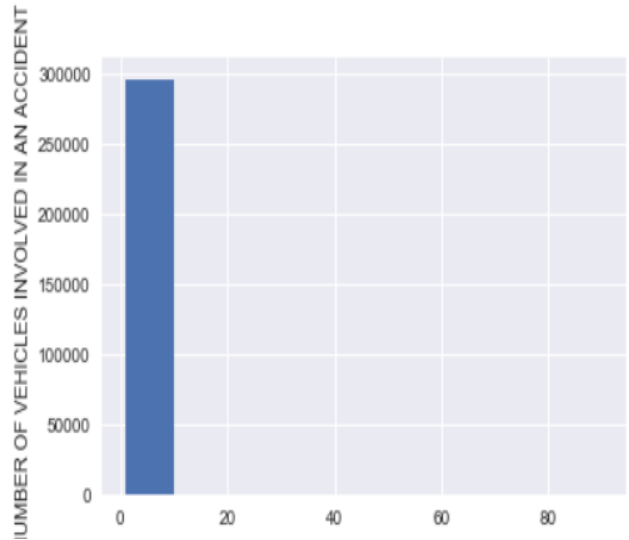
[1 0 2 3 4 5 10 6 15 7 8 9]

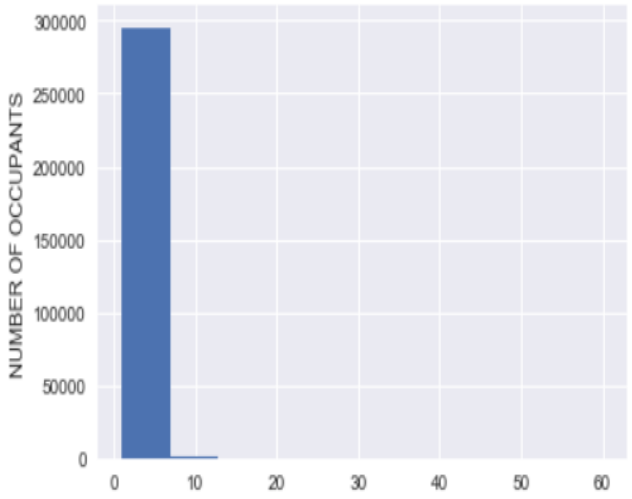


NUMBER PREV SPEED CONV

[0 1 2 5 3 4 6 8]

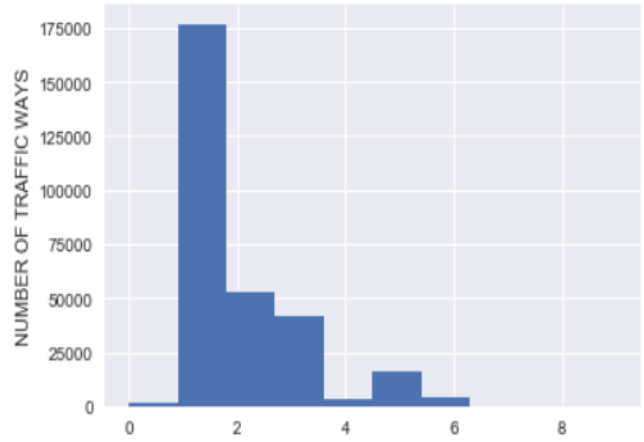




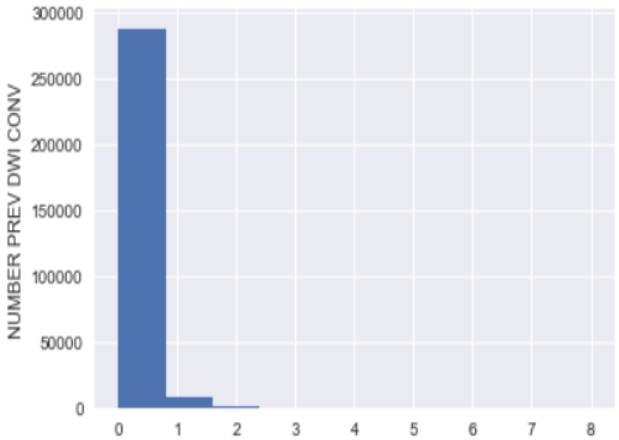


NUMBER OF TRAFFIC WAYS

[1 2 6 5 3 0 4 8 9]



[0 1 2 5 3 4 6 8]



MULTIVARIATE ANALYSIS:

Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle. Multivariate Analysis includes many statistical methods that are designed to allow you to include multiple variables and examine the contribution of each.

PART 1: POISSON REGRESSION

Dimensional Reduction via Feature Selection:

Feature selection is an important part of building machine learning models. This done using two approaches:

1. STATISTICAL RELATIONSHIP:

- CORRELATION

Correlation is any of a broad class of statistical relationships involving dependence. In common usage it most often refers to how close two variables are to having a relationship with each other. In this case, as most of the variables are categorical numerical correlation cannot be implemented. For this, devised an algorithm that works like correlation to give relation between variables. This algorithm checks the data type of each variable works as follows:

1. If the pair of variables are nominal (category) then it performs Cramer's rule on them.
2. If the pair of variables are continuous then it performs Pearson's correlation.
3. If the variable in the pair is either continuous or nominal then correlation measure between a nominal and an numeric is Eta also called as correlation ratio and equal to the root R-square of the one way ANOVA

From the result, decided to drop few variables which showed high correlation values.

- CHI SQUARE TEST

The Chi-Square statistic is most commonly used to evaluate Tests of Independence when using a cross tabulation (also known as a bivariate table). Cross tabulation presents the distributions of two categorical

variables simultaneously, with the intersections of the categories of the variables appearing in the cells of the table. It is used to determine whether there is a significant association between the two variables. The algorithm uses p-values for comparing against the threshold.

Rules to use the Chi-Square Test:

1. Variables are Categorical
2. Frequency is at least 5
3. Variables are sampled independently

If p-value is low then null hypothesis should go. From the results, the values with higher than the threshold are ignored. Thus could eliminated few more variables from the set.

NOTE: A statistically significant result may not be practically significant.

2. REGULARIZED MODEL:

Regularization is a method for adding additional constraints or penalty to a model

This is based on the idea that when all features are on the same scale, the most important features should have the highest coefficients in the model, while features uncorrelated with the output variables should have coefficient values close to zero. Lasso forces weak features to have zero as coefficients.

- LASSO REGULARIZATION AS A FEATURE SELECTION TOOL

Thus L1 regularization produces sparse solutions, inherently performing feature selection. Regularized linear models are a powerful set of tool for feature interpretation and selection. Lasso produces sparse solutions and as such is very useful selecting a strong subset of features for improving model performance.

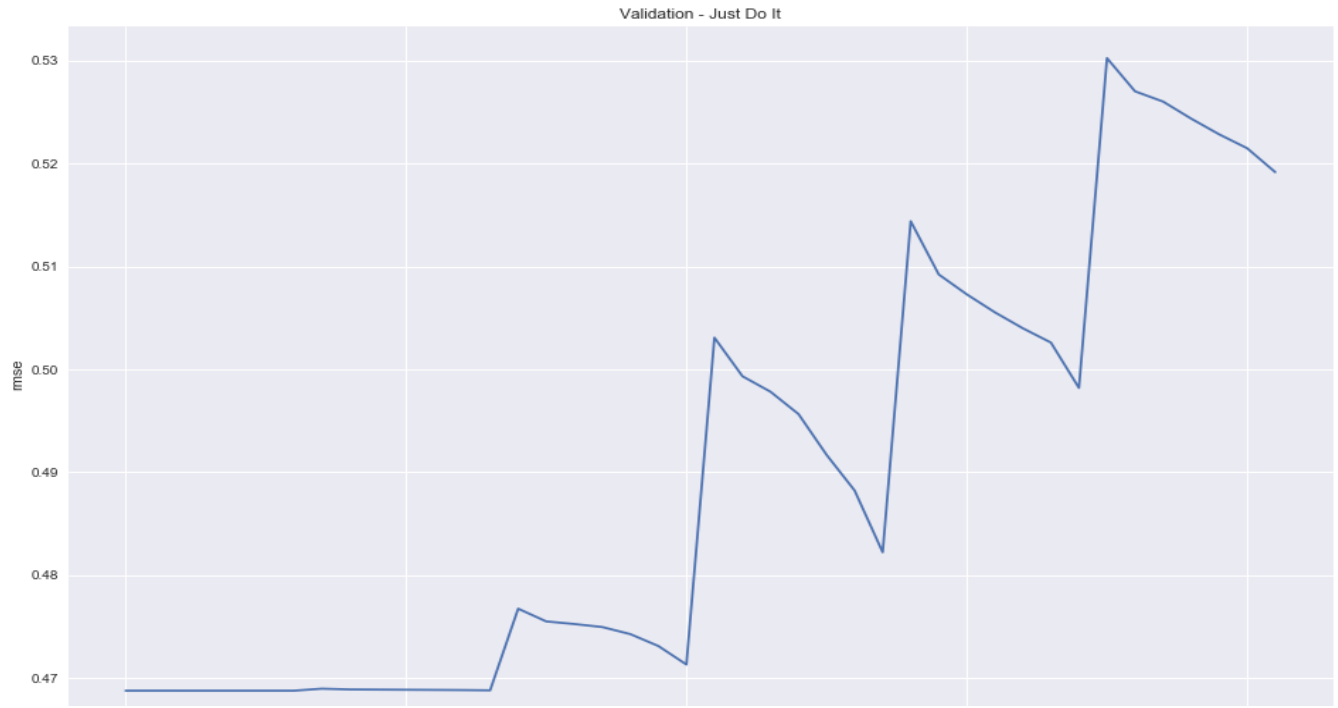
The hyper parameter in Lasso is alpha which has to be tuned. This is done using cross validation and finding an optimal value of the alpha. From the observation, there are 9 features whose coefficient is 0.

- ELASTIC NET AS A FEATURE SELECTION TOOL

Lasso often tends to “over-regularize” a model that might be overly compact and therefore under-predictive. The Elastic Net addresses the aforementioned “over-regularization” by balancing between LASSO and ridge penalties. In particular, a hyper-parameter, namely Alpha, would be used to regularize the model such that the model would become a LASSO in case of Alpha = 1 and a ridge in case of Alpha = 0. In practice, Alpha along with lambda can be tuned easily by the cross-validation. The value selected from cross validation for alpha – lambda is 0.0001 – 0.5.

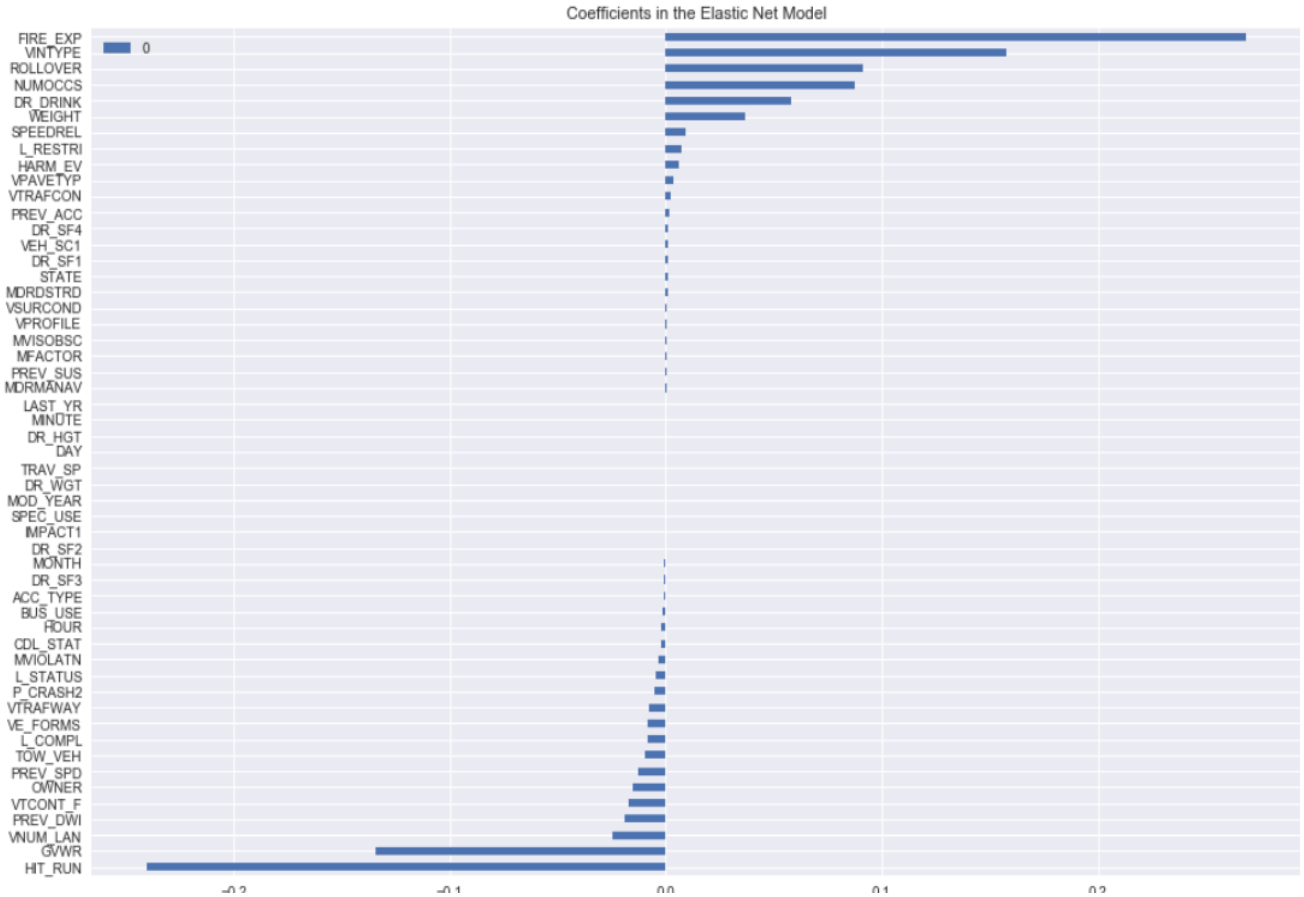
```
#lt.rcParams['figure.figsize'] = (12.0, 6.0)
idx = list(product(alphas, l1_ratios))
p_cv_elastic = pd.Series(cv_elastic, index = idx)
p_cv_elastic.plot(title = "Validation - Just Do It")
plt.xlabel("alpha - l1_ratio")
plt.ylabel("rmse")
```

```
Text(0, 0.5, 'rmse')
```



The performance of Elastic Net was not as expected. The features selected by the model are the same as the features selected by the lasso.

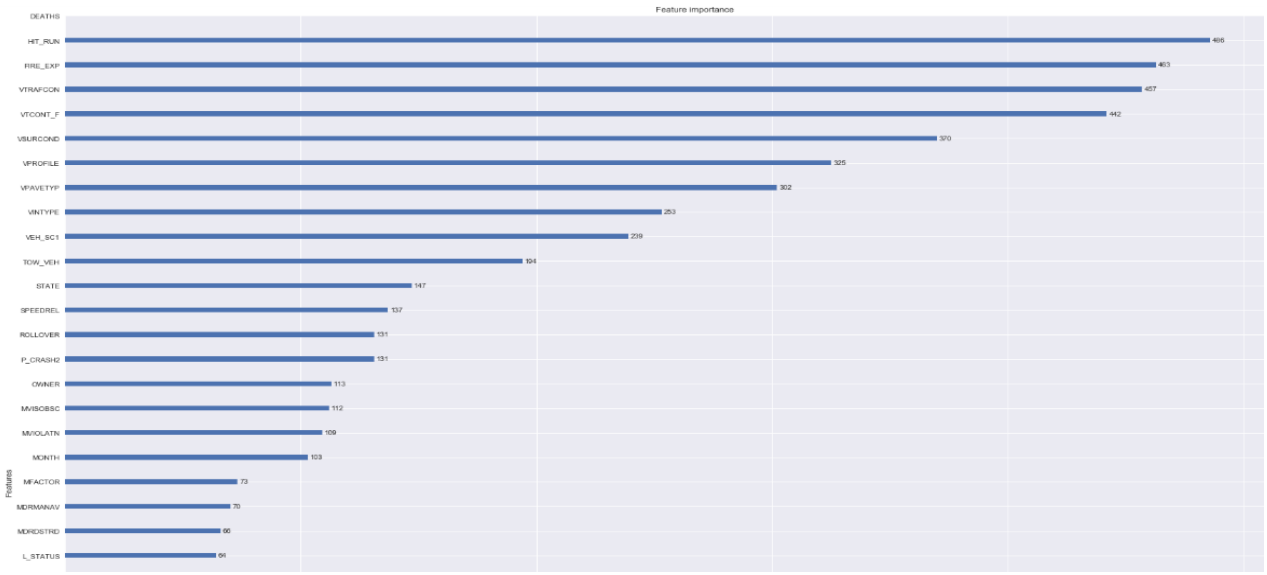
The variable performance scale in elastic net is as follows:



The most important positive feature is FIRE_EXP, then VINTYPE. Other important negative features are HIT_RUN, GVWR, VNUM_LAN etc.

FEATURE IMPORTANCE USING XGBOOST MODEL:

If we look at the feature importance returned by XGBoost below we see that HIT_RUN dominates the other features, clearly standing out as the most important predictor of income. To build the baseline Poisson model, consider all the count variables and top 5 variables from the significant features. The new dataframe contains 14 variables –: 9 count variables and 5 important variables.



The result of the model fitting is as below:

```
po_results = sm.GLM(response, predictors, family=sm.families.Poisson()).fit()
print(po_results.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 DEATHS   No. Observations:               296847
Model:                            GLM   Df Residuals:                   296833
Model Family:                 Poisson   Df Model:                           13
Link Function:                    log   Scale:                            1.0
Method:                          IRLS   Log-Likelihood:            -2.6807e+05
Date:                Mon, 03 Dec 2018   Deviance:                   1.9841e+05
Time:                        03:13:39   Pearson chi2:                 1.13e+09
No. Iterations:                     8
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.1808      0.009     21.039      0.000       0.164       0.198
VSURCOND       0.0017      0.000      5.291      0.000       0.001       0.002
VTRAFCON       0.0052      0.000     19.061      0.000       0.005       0.006
VTCONT_F      -0.0442      0.003    -16.427      0.000      -0.049      -0.039
HIT_RUN       -1.3533      0.027    -49.644      0.000      -1.407      -1.300
FIRE_EXP       0.4385      0.011     40.497      0.000       0.417       0.460
VTRAFWAY      -0.0252      0.002    -11.625      0.000      -0.029      -0.021
NUMOCCS        0.0594      0.001     72.334      0.000       0.058       0.061
VNUM_LAN      -0.0875      0.003    -30.970      0.000      -0.093      -0.082
VE_FORMS      -0.3100      0.003    -92.929      0.000      -0.317      -0.303
PREV_ACC      -0.0107      0.005     -2.233      0.026      -0.020      -0.001
PREV_DWI       0.0801      0.009      8.589      0.000       0.062       0.098
PREV_SUS       0.0157      0.001     12.140      0.000       0.013       0.018
PREV_SPD      -0.0072      0.004     -2.005      0.045      -0.014      -0.000
==============================================================================
```

The Pearson statistic provides an estimate of the data's dispersion. When the data is drawn from a Poisson distribution with sufficient samples the ratio Pearson chi2 / Degree of Freedom for Residuals is approximately 1; for observed data a ratio less than 1 implies underdispersion and more than 1 implies overdispersion. Data that is underdispersed requires a zero-inflated model.

In this case the result, 3806.8, suggests overdispersion, which is amenable to Negative Binomial regression. We explore that below:

```
nb_results = sm.GLM(response, predictors,family=sm.families.NegativeBinomial
                    (alpha=0.671)).fit()
print(nb_results.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                 DEATHS   No. Observations:                296847
Model:                            GLM   Df Residuals:                    296833
Model Family:        NegativeBinomial   Df Model:                           13
Link Function:                    log   Scale:              2045.4528128499728
Method:                          IRLS   Log-Likelihood:            -2.9501e+05
Date:                Mon, 03 Dec 2018   Deviance:                   1.5576e+05
Time:                        03:16:35   Pearson chi2:                 6.07e+08
No. Iterations:                     19
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.0627      0.468      0.134      0.893      -0.855       0.980
VSURCOND       0.0017      0.018      0.096      0.924      -0.033       0.036
VTRAFCON       0.0053      0.015      0.351      0.726      -0.024       0.035
VTCONT_F      -0.0424      0.143     -0.297      0.766      -0.322       0.237
HIT_RUN       -1.3517      1.309     -1.033      0.302      -3.916       1.213
FIRE_EXP       0.4541      0.632      0.719      0.472      -0.784       1.692
VTRAFWAY      -0.0253      0.117     -0.216      0.829      -0.255       0.204
NUMOCCS        0.1151      0.071      1.615      0.106      -0.025       0.255
VNUM_LAN      -0.0849      0.149     -0.569      0.570      -0.377       0.208
VE_FORMS      -0.3003      0.173     -1.739      0.082      -0.639       0.038
PREV_ACC      -0.0056      0.259     -0.022      0.983      -0.514       0.503
PREV_DWI       0.0862      0.528      0.163      0.870      -0.949       1.121
PREV_SUS       0.0159      0.074      0.214      0.831      -0.130       0.162
PREV_SPD      -0.0065      0.195     -0.034      0.973      -0.389       0.376
==============================================================================
```

Again the ratio Pearson chi2 / Degree of Freedom for Residuals is 2021.3 which is too large.

CONCLUSION: The data is highly over dispersed. Statistically speaking, is a Poisson distribution even appropriate? Our model was founded on the belief that the number deaths can be accurately expressed as a Poisson distribution. If that assumption is misguided, then the model outputs will be unreliable.

Thus concluded that the Poisson distribution is not appropriate for the dataset.

PART 2: LOGISTIC REGRESSION

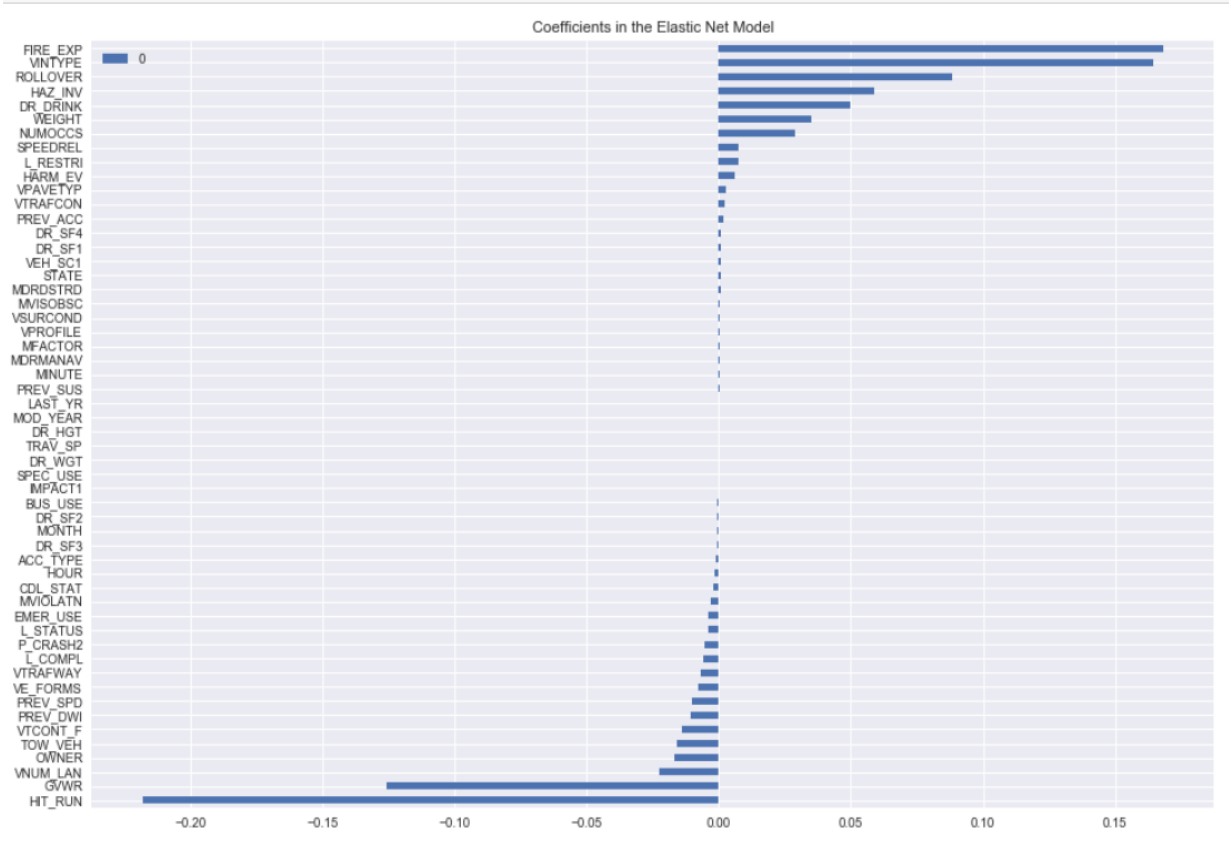
The target variable deaths is now converted to a binary variable.

Dimensional Reduction:

CASE 1:

Some set of steps are implemented again for binary DEATHS variable (DEATHS_BIN). Correlation and Chi square test gives 55 variables. Continuing further with Elastic net for feature selection, the set of variables reduces to 49.

The variable performance scale in elastic net is as follows:



The most important positive feature is FIRE_EXP, then VINTYPE, ROLLOVER. Other important negative features are HIT_RUN, GVWR, VNUM_LAN.

Modeling & Testing:

After fitting a Logistic model to all the 49 variables achieved, the accuracy comes to 82.7%

CASE 2:

STOCHASTIC DIFFUSION SEARCH FOR FEATURE SELECTION:

The method is used to select the most relevant feature subset from the dataset for classification work. The feature selection technique seek to choose a subset of features thereby facilitating performance maximization. Although the approach takes feature dependency into account, one common problem is that it has a higher risk of overfitting than alter techniques, it can be slow to perform and often tends to be computationally intensive.

The algorithm was implemented using two set of estimators:

1. Logistic Regression + Decision Trees
2. Logistic Regression

With the first part, the features selected are:

```
['GVWR', 'WEIGHT', 'BUS_USE', 'DRIMPAIR', 'DR_SF3', 'DR_SF4', 'EMER_USE', 'HO
UR', 'IMPACT1', 'L_COMPL', 'MDRMANAV', 'MONTH', 'MVIOLATN', 'P_CRASH2', 'STAT
E', 'VEH_SC1', 'VINTYPE', 'PREV_DWI', 'NUMOCCS', 'PREV_ACC', 'VNUM_LAN', 'FIR
E_EXP', 'TRAV_SP']
```

With the second part, the features selected are:

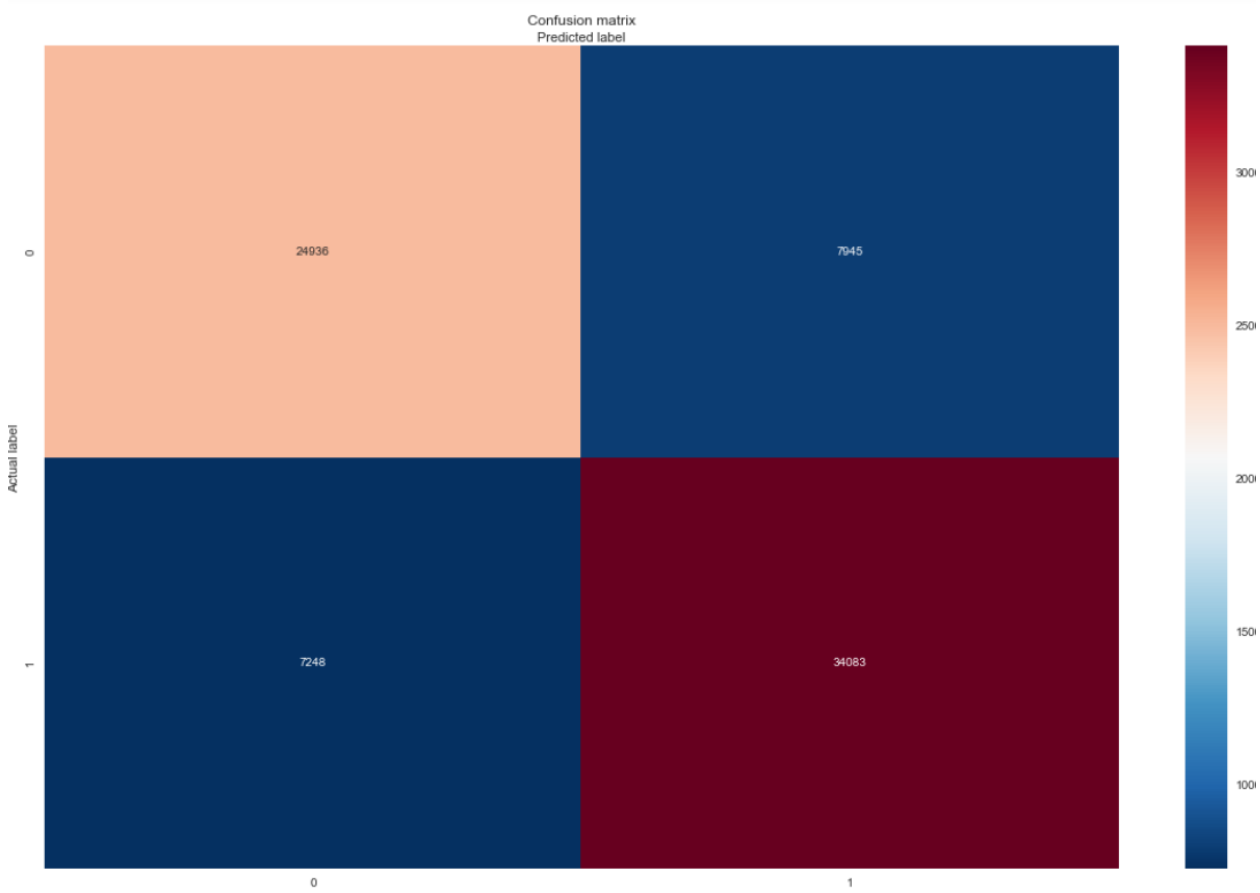
```
['WEIGHT', 'CDL_STAT', 'DR_DRINK', 'LAST_YR', 'MDRDSTRD', 'MOD_YEAR', 'P_C
RASH2', 'ROLLOVER', 'STATE', 'TOW_VEH', 'VEH_SC2', 'VINTYPE', 'VPAVETYP',
'VSURCOND', 'VTCONT_F', 'PREV_DWI', 'NUMOCCS', 'PREV_SPD', 'VNUM_LAN', 'VT
RAFWAY', 'FIRE_EXP', 'HAZ_INV', 'TRAV_SP']
```

Modeling & Testing:

The accuracy score after model fitting is 80.16% and 81.91% for Logistic regression and Decision tree respectively.
The accuracy score in second case for Logistic Regression is 79.52%

Now there is a trade-off here. From the observations, it is noticeable that the second part gives more relevant and reliable set of features compared to the first one even though the performance for the first is better.
So the features from second part are picked for next step of modeling.

The confusion matrix:

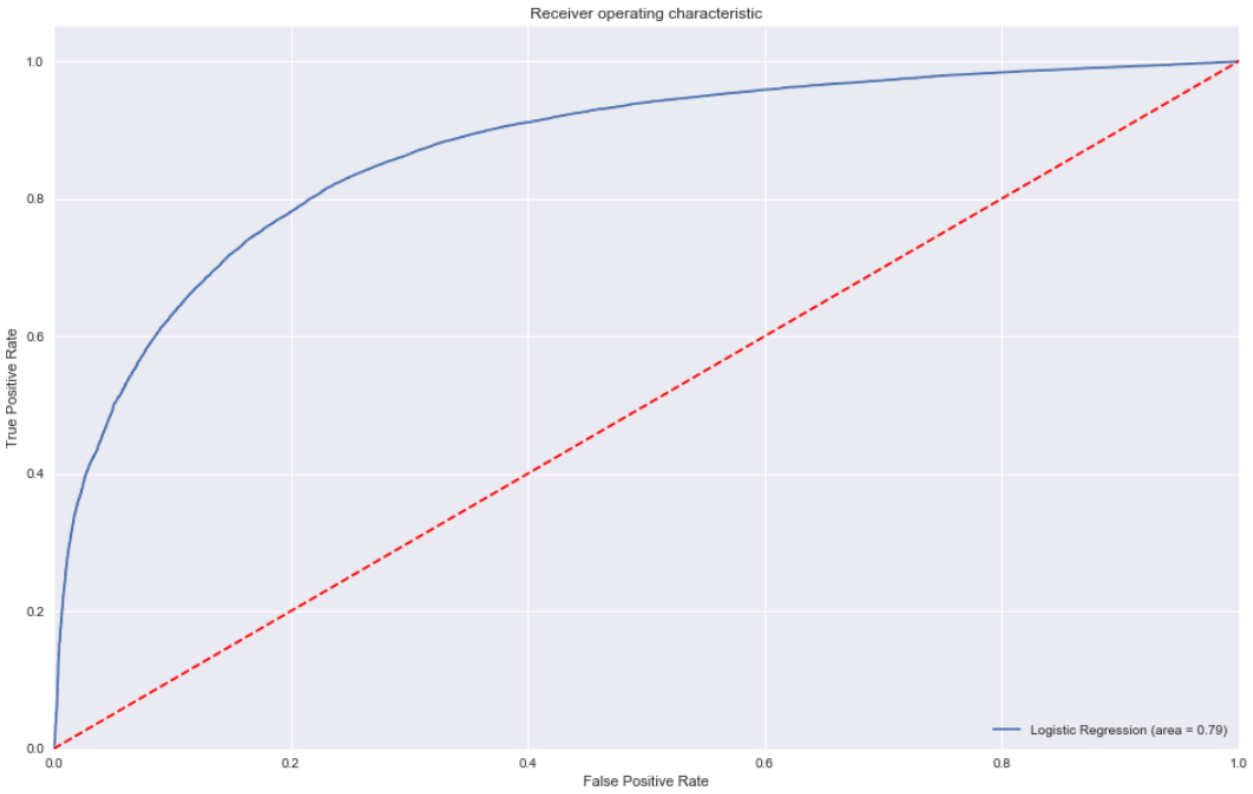


The results are telling us that we have 24936+34083 correct predictions and 7248+7945 incorrect predictions.

The classification report:

	precision	recall	f1-score	support
0	0.77	0.76	0.77	32881
1	0.81	0.82	0.82	41331
avg / total	0.79	0.80	0.80	74212

Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate against the false positive rate. It shows the tradeoff between sensitivity and specificity. The ROC score is 0.79.



AUC score for the case is 0.8695. AUC score 1 represents perfect classifier, and 0.5 represents a worthless classifier.

<u>INFERENCE FROM THE 2 CASES ABOVE:</u>
From case one we have set of 49 features with model accuracy as ~83% while on the other side we have set of 23 features with model accuracy as 79.5.
Accuracy concerns the ability of a model to make correct predictions, while interpretability concerns to what degree the model allows for human understanding. Models exhibiting the former property are many times more complex and opaque, while interpretable models may lack the necessary accuracy.

The trade-off between accuracy and interpretability for predictive modeling has to be handled. Thus the model with fewer parameters is easier to interpret.
Hence the 23 feature set is selected for model interpretation

INTERPRETATIONS:

The interpretations from the coefficients is done through the concept of odds ratio and its probability and try to understand the logistic regression results.

FIRE_EXP, HAZ_INV, WEIGHT

	Coefficients	Odds_ratio	Probability
Intercept	0.182119	1.199757	54.540442
FIRE_EXP	1.588015	4.894026	83.033667
HAZ_INV	0.227126	1.254988	55.653876
TRAV_SP	0.000305	1.000305	50.007636
WEIGHT_1	0.735569	2.086669	67.602621
WEIGHT_2	0.060229	1.062080	51.505266
WEIGHT_3	-0.389014	0.677725	40.395478
WEIGHT_4	-0.129883	0.878198	46.757476
WEIGHT_5	-0.126789	0.880920	46.834526
WEIGHT_6	-0.303486	0.738241	42.470562
WEIGHT_7	-0.281906	0.754344	42.998647
WEIGHT_8	-0.611663	0.542448	35.168002
WEIGHT_9	1.229061	3.418019	77.365423

The fitted model says that, holding all the variables at a fixed value, the odds of deaths when there is fire explosion over the odds of deaths when there is no fire explosion is 4.89. In terms of percentage change, we can say that the odds of fire explosions are 83.03% higher than the odds of no fire explosion.

Similarly, the odds of having hazardous placards in the vehicle are 55% higher than the odds of non-involvement of hazardous placards.

In case of Weight of the vehicle, the category WEIGHT_1 and WEIGHT_9 represent light vehicle like cars & 4 wheelers that are 6000lbs or less and unknowns which often include buses.

Thus holding all the variables at a fixed value, the percentage of odds of light vehicles is 67% higher than heavy vehicles. On the same path, holding all the variables at a fixed value, the percentage of odds of buses is 77% higher than other vehicles.

DR_DRINK

DR_DRINK_0	0.000135	1.000135	50.003381
DR_DRINK_1	0.181984	1.199595	54.537089

The interpretation for whether the driver was drinking seems justifying. The odds of drunk drivers is 1.19 over the odds of non-drunk drivers. That is odds of drunk drivers are 55% higher than the odds of non-drunk drivers.

P_CRASH2

P_CRASH2_1	0.491506	1.634777	62.046118
P_CRASH2_2	-0.001150	0.998850	49.971245
P_CRASH2_3	0.079225	1.082447	51.979577
P_CRASH2_4	0.013443	1.013534	50.336064
P_CRASH2_5	0.671685	1.957534	66.188046
P_CRASH2_6	2.178651	8.834382	89.831593
P_CRASH2_8	0.991419	2.695057	72.936817
P_CRASH2_9	1.149812	3.157601	75.947666
P_CRASH2_10	0.716509	2.047274	67.183780
P_CRASH2_11	0.530000	1.698933	62.948318
P_CRASH2_12	3.319180	27.637680	96.508097
P_CRASH2_13	2.828750	16.924291	94.420979
P_CRASH2_14	0.581706	1.789088	64.145992

This attribute defines the attribute that best describes the critical event which made the collision possible. Categories from 1 to 9 represent vehicle's loss of control while categories 10 to 14 represent vehicle travelling direction and these two categories are the dominant death causing factor in P_CRASH2.

ROLLOVER

ROLLOVER_0	-1.307623	0.270462	21.288487
ROLLOVER_1	0.167552	1.182407	54.179023
ROLLOVER_2	1.132543	3.103538	75.630787
ROLLOVER_9	0.189648	1.208824	54.727034

The probability percentage for odds of vehicle involved in rollover or overturn is 76% which is significantly higher. The model says that when there is a case of rollover death/s is involved.

VEH_SC2

VEH_SC2_0	0.168667	1.183726	54.206704
VEH_SC2_30	0.000000	1.000000	50.000000
VEH_SC2_32	0.001504	1.001505	50.037593
VEH_SC2_35	0.000116	1.000116	50.002900
VEH_SC2_36	-0.000653	0.999347	49.983666
VEH_SC2_44	0.006008	1.006026	50.150210

This attribute defines factors that are related to the vehicle specifically. Some of the levels are multi wheeled vehicles, vehicles for handicaps, altered vehicles or adaptive equipment.

PREV_DWI

PREV_DWI_0	0.179376	1.196471	54.472416
PREV_DWI_1	0.030094	1.030552	50.752303
PREV_DWI_2	0.010528	1.010584	50.263206
PREV_DWI_3	-0.017328	0.982821	49.566802
PREV_DWI_4	-0.015720	0.984402	49.606997
PREV_DWI_5	-0.000457	0.999543	49.988582
PREV_DWI_6	-0.001595	0.998407	49.960132
PREV_DWI_8	-0.002779	0.997225	49.930521

This element records any previous DWI convictions for the driver that occurred with past 5* years. The odds ratio for all the categories is around 1 which can be interpreted as not so significantly contributing factor.

NUMOCCS _

NUMOCCS_1	-0.605269	0.545928	35.313920
NUMOCCS_2	-0.118474	0.888275	47.041615
NUMOCCS_3	0.081894	1.085340	52.046195
NUMOCCS_4	0.210442	1.234223	55.241720
NUMOCCS_5	0.264308	1.302529	56.569501
NUMOCCS_6	0.165849	1.180395	54.136747
NUMOCCS_7	0.065201	1.067374	51.629447
NUMOCCS_8	0.058700	1.060457	51.467079
NUMOCCS_9	0.019303	1.019491	50.482570
NUMOCCS_10	0.011552	1.011619	50.288785
NUMOCCS_11	0.008575	1.008612	50.214383
NUMOCCS_12	-0.001409	0.998592	49.964769
NUMOCCS_13	0.003011	1.003016	50.075284

The fitted model says that, holding all the variables at a fixed value, the odds of deaths when number of occupants is 1 (only driver) over the odds of deaths when the occupants are more is 0.54. In terms of percentage change, we can say that the odds of number of occupants is 1 are 35.3% lower than the odds of more number of occupants. Thus there is less chances of people dying when the occupants are less which logically makes sense.

VNUM_LAN

VNUM_LAN_0	0.006637	1.006659	50.165916
VNUM_LAN_1	0.228396	1.256583	55.685206
VNUM_LAN_2	0.386010	1.471099	59.532172
VNUM_LAN_3	0.041152	1.042011	51.028659
VNUM_LAN_4	-0.097193	0.907381	47.572094
VNUM_LAN_5	-0.117005	0.889580	47.078197
VNUM_LAN_6	-0.182872	0.832875	45.440907
VNUM_LAN_7	-0.083005	0.920346	47.926061

This attribute describes number of travel lanes the vehicle driving just before the accident. From the odds ratio, it’s clear that the odds of deaths is less frequent when the number of lanes are more.

VINTYPE

VINTYPE_0	-1.191855	0.303657	23.292736
VINTYPE_1	-1.160780	0.313242	23.852552
VINTYPE_2	2.617835	13.706022	93.200065
VINTYPE_3	-0.079402	0.923669	48.015992
VINTYPE_4	-0.003679	0.996328	49.908033

The levels are encoded as VINTYPE_0: Passenger vehicle, VINTYPE_1: Truck, VINTYPE_2: Motorcycle, VINTYPE_3: Unknown, VINTYPE_4: Commodity vehicles. We can see the odds of motorcycle riders is significantly higher than the odds of other vehicle. Thus motorist are more prone to deaths than others.

VTRAFWAY

VTRAFWAY_0	0.006637	1.006659	50.165916
VTRAFWAY_1	0.023935	1.024224	50.598353
VTRAFWAY_2	0.020755	1.020972	50.518859
VTRAFWAY_3	0.094472	1.099079	52.360050
VTRAFWAY_4	-0.146965	0.863324	46.332472
VTRAFWAY_5	-0.085122	0.918400	47.873226
VTRAFWAY_6	0.269176	1.308886	56.689070
VTRAFWAY_8	0.003060	1.003065	50.076511
VTRAFWAY_9	-0.003829	0.996178	49.904267

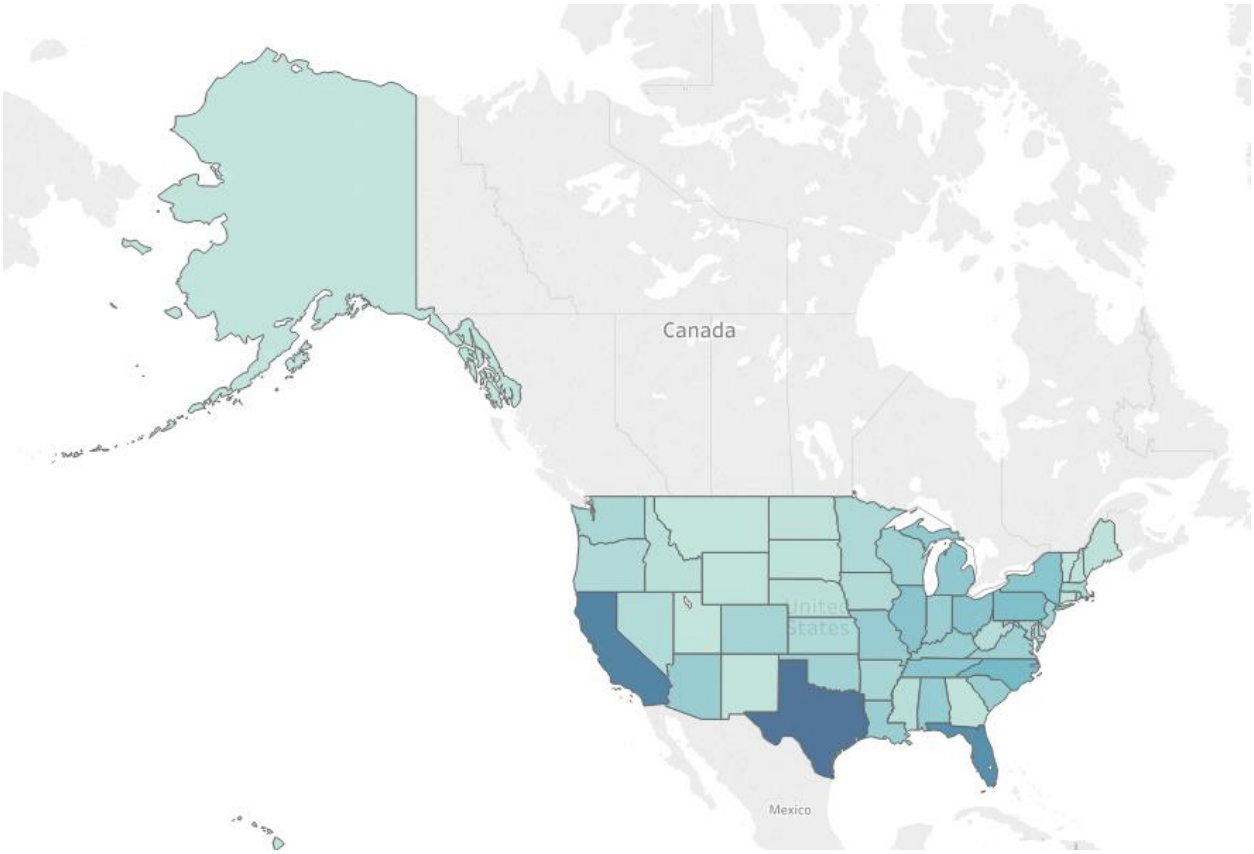
This attribute describes the traffic flow just before the vehicle's precrash event. The probability of all the categories contributing in the same -50%. But little dominantly we can say VTRAFWAY_6 which represents entrance/exit ramp have higher odds ratio than the rest.

VTCONT_F

VTCONT_F_0	-0.135982	0.872858	46.605673
VTCONT_F_1	0.007486	1.007514	50.187157
VTCONT_F_2	-0.011842	0.988227	49.703943
VTCONT_F_3	0.271845	1.312384	56.754581
VTCONT_F_8	0.028635	1.029049	50.715819
VTCONT_F_9	0.021978	1.022221	50.549427

This variable identifies traffic control device functioning. The odds ratio is higher for VTCONT_F_3 (1.312) which represents traffic devices functioning properly compared to others.

STATES



The number of accident resulting in deaths seems really high California, Texas followed by Florida, North Carolina.

CHALLENGES:

- Data integration was a hectic task. As there were 20 different files available for each year which comes to around 160 files for 8 years. Also some of the files were not at all in sync with files across years so horizontal merge (merging across years) was tricky and messy.
- Deciding the target variable
- Dimensionality reduction: Implemented many approaches that were relevant to the structure of the data and its variable. Deciding and finalizing the ones which were suitable.
- Analyzing and making sense of the variables that are given out by the model and comprehending it into meaning results.