PA Model Solution June 14, 2019

This model solution is provided so that candidates may better prepare for future sittings of Exam PA. It includes both a sample solution, in plain text, and commentary from those grading the exam, in italics.

In many cases there is a range of fully satisfactory approaches. This solution presents one such approach, with commentary on some alternatives, but there are valid alternatives not discussed here.

Exam PA June 2019 Project Report Template

Instructions to Candidates: Please remember to avoid using your own name within this document or when naming your file. There is no limit on page count.

As indicated in the instructions, work on each task should be presented in the designated section for that task.

Task 1 – Explore the relationship of each variable to *Crash_Score* (5 points)

The best candidates reached conclusions based on both graphs and summary statistics. It was not necessary to provide supporting information for each variable. The key was to demonstrate the ability to draw inferences from the graphs and tables.

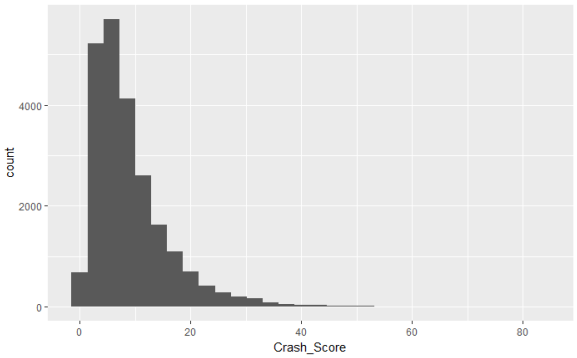
Some candidates spent too much time on this task. At five points it should translate to about 15 minutes of work.

Many candidates failed to consider the skewed distribution of the target variable before launching into the analysis. It is a common practice to make a transformation when the target variable is skewed. Candidates who did not make the transformation were not further penalized for making inferences based on the original values of Crash Score.

Because there is no clear boundary between appearing to be predictive and not, it was not necessary for candidates to reproduce the same list as presented here.

It was important to point out the nature of the relationship to the target variable.

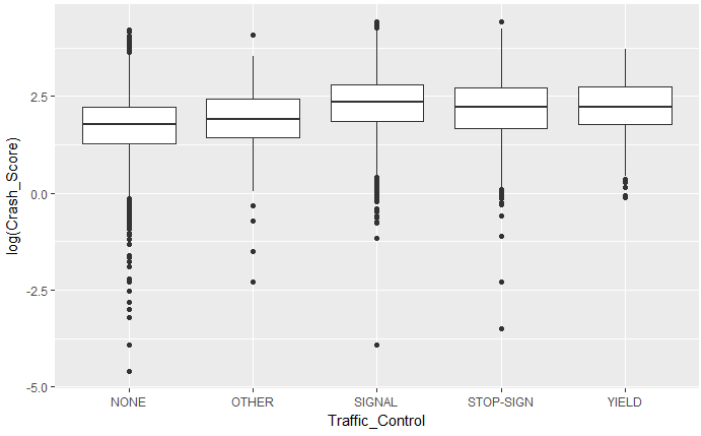
For the target variable Crash Score, the median is 7.16, the mean is 9.11, and the maximum is 83.41. This indicates that the distribution is skewed to the right. A histogram confirms this:



As a result, I explored boxplots of the log of the target variable split among the factors of each variable. Differences were observed for the following variables:

- Time_of_Day: Low Crash Score for period 1 (midnight to 4am)
- Rd_Feature: High for INTERSECTION and RAMP
- Rd_Character: High for OTHER
- Rd_Configuration: High for TWO-WAY-UNDPROTECTED-MEDIAN
- Rd_Surface: Low for OTHER
- Rd_Conditions: High for WET, low for OTHER
- Light: Low for DARK-NOT-LIT and OTHER
- Weather: Low for OTHER
- Traffic_Control: Low for NONE and OTHER
- Work_Area: Low for NO.

The plot for Traffic_Control is provided below. The others can be obtained from the R code.



Looking at means and medians for the logarithms of crash scores reveals some other possible relationships beyond those already mentioned:

- Month: Higher in months October (10) through March (3)
- Rd_Class: Higher for US HWY
- Rd_Surface: Also, higher for the two ASPHALT levels relative to the two CONCRETE levels

An example using Rd_Surface appears below.

Rd_Surface <fctr>	mean <dbl>	median <dbl>	n <int>
SMOOTH ASPHALT	1.946158	1.986504	20007
COARSE ASPHALT	1.913299	1.931521	1997
CONCRETE	1.688533	1.704746	692
GROOVED CONCRETE	1.708268	1.796747	371
OTHER	1.450069	1.611677	70

It appears there are several variables that may predict the target variable, but it should also be noted that none stand out as making large differences.

Task 2 – Reduce the number of factor levels where appropriate (5 points)

Candidates generally did well on this task. The best candidates found multiple cases where factor levels could be combined, considered the similarity of means/medians, and provided adequate rationale for combining levels. It was not sufficient to only combine those levels with extremely low counts.

The number of variables for which factors were combined was less important than the quality of the combinations made and the supporting evidence.

The following combinations are made:

- Time_of_Day: Time 1 = OVERNIGHT, Times 2 and 6 = LATE-EARLY, Times 3-5 = DAYTIME. They have different means and medians and make sense with regard to accident severity.
- Rd_Feature: Combine INTERSECTION and RAMP into one level, INTERSECTION-RAMP, and combine the others into OTHER. Intersection and ramp accidents are more likely to involve multiple vehicles and hence more damage.
- Rd_Character: Based on differing mean scores, combine the three with OTHER into OTHER and combine the remaining levels as STRAIGHT or CURVED.
- Traffic_Control: Combine NONE and OTHER into OTHER and the others into CONTROLLED to reflect some sort of control.

The other predictor variables either show little difference between the factor levels or enough differences throughout so that no obvious groupings exist.

Task 3 – Use observations from principal components analysis (PCA) to generate a new feature (9 points)

Candidates needed to consider the variance explained and the relative coefficients within loadings. Many candidates misinterpreted the loadings or did not know how to use them to create a new feature. For example, some candidates used only the large positive or only the large negative coefficients. Other common mistakes were to use a component other than PC1 or to use a combination of PCs (since each PC is the best available linear combination there is no reason to combine them).

Using every number in the first principal component is not as beneficial as it will be difficult to explain to the client what it represents. By using only a few of the factor levels it is clearer what the new variable is measuring.

Many candidates did not drop the contributing variables to the new feature, leading to a rank-deficient model in Task 5.

To receive full credit, the candidate had to create a new feature based on the PCA.

Running the PCA on these three variables shows that only 22% of variation is explained by the first principal component (PC) and 35% by the first two PCs. However, the loadings may highlight interesting relationships among these variables.

The largest loadings on the first PC are:

- Rd_ConditionsDRY: -0.51
- Rd_ConditionsWET: 0.50
- WeatherCLEAR: -0.46
- Weather RAIN: 0.43

Applying these weights creates a variable that is strongly positive for rain/wet conditions and strongly negative for dry/clear conditions. It makes sense to pair up each of these as they would typically appear together, e.g. rain leads to wet roads.

Based on these results, I created a new feature, WETorDRY, based on the Rd_Conditions and Weather variables, deleting these two but retaining the Light variable as is.

Task 4 –Select an interaction (7 points)

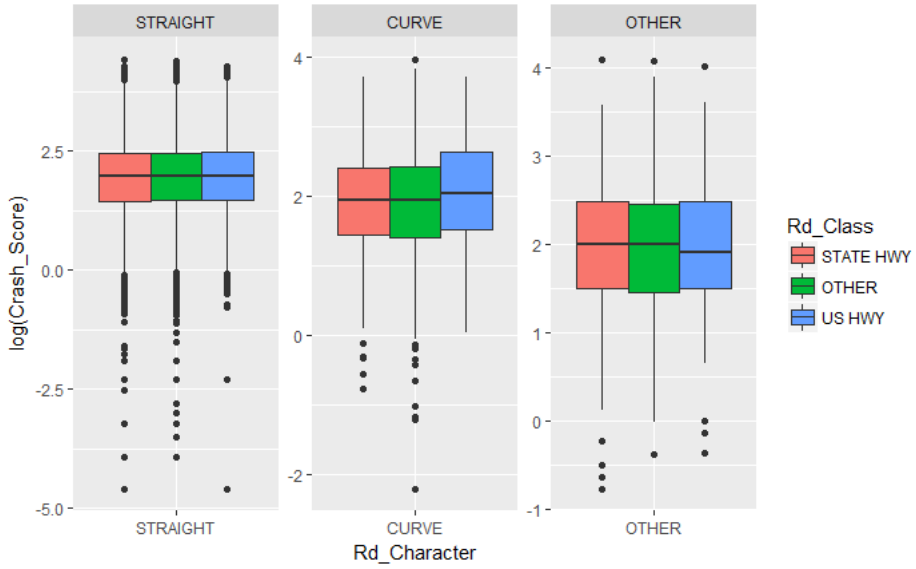
The best candidates began by explaining what they were looking for when searching for an interaction. They next proposed an interaction based on an understanding of traffic behavior. They then used a graph to either confirm or dispel their interaction. In some cases, it was then necessary to try a new pair of variables.

Candidates were not required to create the interaction variable at this task. However, points were lost if the selected interaction was not used in subsequent tasks.

While the plots presented here used the logarithm of Crash_Score, it was not necessary to do so to earn full credit.

An interaction is indicated when changing the level of one variable alters how levels of the other variables affect the target.

A first thought is Rd_Character and Rd_Class. Changing Rd_Character from STRAIGHT to CURVE may have a different effect depending on the Rd_Class. U.S. highways may have gentler curves than state highways and hence a different effect. As seen below, there does seem to be an interaction. For the CURVE cases the crash score is higher for US HWY. While this is contrary to my intuition, the interaction appears to be worth considering.



I'll use this one for future work.

Task 5 – Select a distribution and link function (10 points)

Many candidates noted that the target variable is a positive, continuous variable with a right skew. Strong candidates translated this understanding into their model choices.

Many candidates went with the canonical link function just because it is canonical. That is a weak justification. The inverse link can make negative predictions. The inverse square link function does make positive predictions but is also hard to interpret. Ability to interpret the predictions will be important for the client and this should inform the decision made. A better choice for the link function is the log link. It ensures positive predictions and is easy to interpret.

The best choices for distribution are gamma and inverse Gaussian as both are positive and right skewed. The Gaussian distribution admits negative observations. It was an acceptable choice provided candidates noted this possibility. It turns out that the inverse Gaussian distribution and log link will not converge.

Candidates didn't always incorporate their work from Tasks 2-4 in their model. As noted in Task 3, if a new variable is created and the contributing variables not deleted, the regression will produce a rank-deficient model. Many candidates either did not notice this (or if they did, did not comment on it or were unable to explain it).

Many candidates failed to adequately justify their choices for distribution, link function, and best model. When log link function is applied, the log transformation of the target variable is not necessary.

Before building a GLM, I split the data into training (75%) and testing (25%) sets. The average target value was 9.108 for the training set and 9.109 for the testing set, so the built-in stratification of the target variable worked well.

The variables Month and Year need to be considered no later than this stage. Some candidates had already eliminated them, which is acceptable. Treating Year as a numeric variable makes

some sense. Even though year as a predictor may not be interesting, if there is a trend over time, including this variable will allow the other effects to be more accurately evaluated. Because any month effect is likely to be seasonal, if retained, this variable should be treated as a factor variable.

I have retained Year as a numeric variable in case there is a trend effect that needs to be accounted for. I converted Month to a factor variable so that any seasonal effect can be determined.

The task instructions did not specifically state to run the OLS model. While no points were lost for failing to do so, since the code was available, it is a good idea to run this as a baseline model.

Before investigating various GLMs, an OLS regression was run using the variables previously created but not the interaction from the previous task. Key values were an AIC of 113,388 and an RMSE of 6.2481 on the testing set. This provides a benchmark for further model development.

The only link function I will consider is the log link. This link ensures that all predictions are positive values, which is a characteristic of the target variable. The log link is also easy to interpret.

Because the target variable is highly right skewed, a skewed distribution such as the gamma seems appropriate. For the gamma distribution with the log link and the interaction term, the AIC is 102,336 and the RMSE is 6.2278. Both are improvements over the OLS model in that smaller values are preferred.

Another skewed distribution is the inverse Gaussian. Running it with the log link did not converge, so this combination could not be evaluated.

A Gaussian distribution with the log link, while not right skewed, will ensure positive predictions. For the normal distribution with the log link and the interaction term, the AIC is 113,208 and the RMSE is 6.2390.

Based on these numerical results (the gamma distribution has a lower AIC and essentially the same RMSE), the gamma distribution with the log link will be used from here on. Given that the two models have the same number of parameters, the lower AIC indicates that the loglikelihood is larger and thus the model fits better to the training data.

Task 6 – Select features using AIC or BIC (12 points)

Many candidates failed to accurately describe AIC, BIC, forward selection, and backward selection, including how they work and why someone would use them. The best candidates carefully considered which criterion (AIC vs BIC) and which selection process (forward vs backward) to use and were able to adequately justify their choices based on the context of the problem. Any combination of criterion and selection process is valid if justified appropriately.

Being more comfortable with one method, a method being the default, or a method being commonly used are not satisfactory justifications for selecting a method.

Some candidates were confused about the nature of penalty imposed by AIC and BIC. The fact that BIC employs sample size does not imply that the actual sample size should play a role in deciding which method to use.

Some candidates used BIC for variable selection and then noted that the AIC was not minimized. This is to be expected and does not imply that BIC was a poor choice.

Some candidates binarized the factor variables so that the stepwise selection could work with individual factor levels. This was not required and candidates did not gain or lose points for doing so. However, candidates who did not binarize should note that some retained factor levels are not significant.

When a regression model is constructed using a large set of predictor variables there is a risk of overfitting. While additional variables can only improve the fit to the training data, they may actually decrease the fit against unseen (testing) data. One of the methods for handling this is the use of penalized likelihood, also known as information criteria. When fitting models by maximum likelihood, additional variables never decrease the loglikelihood value. An information criterion demands that for an additional variable to be included it must not just increase the loglikelihood, it must do so by at least a specific amount. Two popular criteria are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). For AIC, adding a variable requires an increase in the loglikelihood of two per parameter added. For BIC, the required per parameter increase is the logarithm of the number of observations. For the training dataset it is log(17,354) = 9.76 per parameter.

For this problem, BIC is a more conservative approach as there is a greater penalty for each parameter added, requiring more evidence to support additional variables. Our goal in this project is identify the key variables that relate to the target variable. As such, it makes sense to take a conservative approach and work with as few variables as necessary. Thus, BIC makes the most sense for this analysis.

An argument can also be made for having more variables, to give the client more factors to consider. Either answer can earn full credit provided the candidate understood how the choice of AIC and BIC relates to the number of variables selected. The same comment applies to forward versus backward selection.

Similarly, forward selection is more likely to end up with fewer variables. With forward selection, you start with no variables and then add variables until there is no improvement by the selected criterion. Backward selection starts with all the variables and sequentially removes them until no improvement results. It seems more likely that forward selection will result in a simpler model and hence that approach will be used.

When employing BIC and forward selection, the final model uses the following four features:

- Rd_Feature
- Rd_Configuration
- Time_of_Day
- Traffic_Control

Candidates should run the final model and check that everything looks right. This model is also run in Task 7 and if comments about the model were made as part of validation, that was acceptable.

When running the model with these four variables I noticed that Rd_Configuration = UNKNOWN was not significant compared to the base class. Forward selection with factor variables does not consider

individual factor levels, so this is a possible outcome of this method. Options available include combining this level with the base level or binarizing this variable and rerunning the analysis. Although this level is clearly not significant (p = 0.96), I'm not sure what to do with it as folding it into the base class may be hard to explain. There are only 57 records with this value, so leaving it in should not affect other conclusions.

Task 7 – Validate the model (9 points)

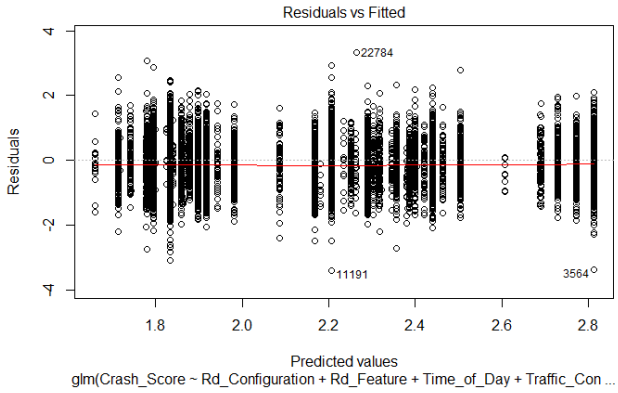
To receive full points, candidates had to produce diagnostic plots and interpret them. After producing the diagnostic plots, many candidates failed to identify places where the plots indicated a problem with the model fit, such as the tails of the q-q plot. It was expected that the diagnostic plots would be used to check model assumptions such as heteroscedasticity and normality of residuals. Some candidates only produced a plot of predicted results versus actuals, which is less helpful than viewing residuals directly.

Some candidates did not understand that the Q-Q plot checks the normality of the standardized deviance residuals. If an appropriate model is being used, this should be the case regardless of the distribution used in the model.

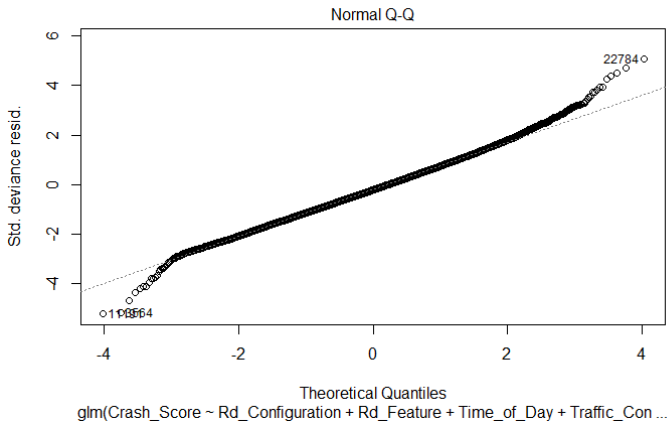
Because a good model was used in this solution, the graphs look reasonably good. Candidates who used other models were likely to get graphs that did not support their model. Full credit was awarded if this conclusion was made.

Running the model with the four features identified in Task 6 produced an RMSE of 6.2301 on the test data. Training and testing the assistant's OLS model on the same data produced an RMSE of 6.2481. The RMSE on unseen data for the GLM is slightly lower despite having far fewer features, suggesting that the GLM is the better model.

The following plot of residuals versus fitted values shows that the model performs well. Because all the predictors are factor variables there are only 2x5x3x2 = 60 possible predicted values, which explains the vertical array. All the vertical bars are centered near zero and spread symmetrically in each direction, indicating constant variance and near zero mean for residuals.



The q-q plot shows that the normal distribution assumption (for the residuals) is maintained for most values, but not near the extremes. It appears a fatter-tailed model may do better.



Task 8 – Interpret the model (9 points)

A few candidates used the model developed on training data coefficients, rather than correctly re-estimate the coefficients using all the data.

Those with inverse or inverse square link functions often had more trouble with interpretation and often got the direction wrong.

For some candidates, it was difficult to verify that the correct model was used in the Rmd code. Sometimes the output could not be found from models found in code.

Many candidates did not interpret or compare the effect of the various coefficients. Merely stating that the direction is toward higher or lower crashes is not sufficient.

Some candidates did not recognize that factors were Boolean and the coefficients do not represent the effect of a 1% increase in the predictor.

It was also expected that candidates would explain the effects in a way that could be easily understood by someone who is not familiar with predictive analytics. The most successful candidates were able to provide a narrative explaining a possible cause and effect for the most important effects.

Some candidates provided model output as an appendix, which is acceptable.

The gamma model with a log link and four features was fit to the full dataset. The results are:

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               1.838598   0.006812 269.913  < 2e-16 ***
Rd_FeatureINTERSECTION-RAMP               0.533185   0.012282  43.411  < 2e-16 ***
Rd_ConfigurationONE-WAY                  -0.059694   0.017914  -3.332 0.000863 ***
Rd_ConfigurationTWO-WAY-PROTECTED-MEDIAN  0.055913   0.014045   3.981 6.89e-05 ***
Rd_ConfigurationTWO-WAY-UNPROTECTED-MEDIAN 0.366560  0.010399  35.251  < 2e-16 ***
Rd_ConfigurationUNKNOWN                   0.016293   0.087165   0.187 0.851719
Time_of_DayOVERNIGHT                     -0.112799   0.023406  -4.819 1.45e-06 ***
Time_of_DayLATE-EARLY                    -0.042275   0.011383  -3.714 0.000205 ***
Traffic_ControlCONTROLLED                 0.072497   0.011890   6.097 1.10e-09 ***
```

Due to the use of the log link, an appropriate way to interpret coefficients is to exponentiate them and subtract 1. The following table provides that interpretation:

Feature	Coefficient	Interpretation
Road Feature = INTERSECTION-RAMP	0.533	70% increase in Crash Score compared to non-intersection-ramp. Crashes at intersections and ramps are likely to involve multiple cars, resulting in higher crash scores.
Rd_Configuration = ONE-WAY	-0.060	6% decrease in Crash Score compared to TWO-WAY-NO-MEDIAN. Less chance of a head-on collision.
Rd_Configuration = TWO-WAY-PROTECTED-MEDIAN	0.056	6% increase in Crash Score compared to TWO-WAY-NO-MEDIAN. This seems odd as protection should minimize head-on collisions, but perhaps overall speeds are higher.
Rd_Configuration = TWO-WAY-UNPROTECTED-MEDIAN	0.367	44% increase in Crash Score compared to TWO-WAY-NO-MEDIAN. This seems odd as a median should reduce head-on collisions, but perhaps overall speeds are higher.
Rd_Configuration = UNKNOWN	0.016	2% increase in Crash Score compared to TWO-WAY-NO-MEDIAN.
Time of Day = OVERNIGHT	-0.113	11% decrease in Crash Score versus daytime (8am to 8pm). It is possible that drivers might be more cautious at night and there are fewer cars on the road.
Time of Day = LATE-EARLY	-0.042	4% decrease in Crash Score versus daytime (8am to 8pm). LATE-EARLY includes times from 4AM to 8AM and 8PM to 12AM. During these times, there are likely fewer cars on the road compared to daytime, leading to lower crash scores.
Traffic Control = CONTROLLED	0.072	7% increase in Crash Score versus no control. Traffic controls tend to be used in areas where there is a lot of traffic. Crashes in these areas are likely to involve multiple vehicles.

One would expect higher crash scores for crashes associated with multiple vehicles and higher speeds. The model output generally aligns with that intuition, however, the results for Rd_Configuration are not intuitive without more subject matter expertise.

Task 9 – Investigate ridge and LASSO regressions (12 points)

The best candidates recognized LASSO and ridge regression as alternatives to forward or backward selection for reducing model complexity. Most recognized that LASSO and ridge regression used a penalty function, although not all recognized that the penalty function was used to prevent overfitting. Most stated that LASSO performed model selection while ridge did not, and some used this idea to recommend LASSO regression from a practical standpoint, as variable selection was identified in the problem statement as a project goal. Candidates who simply pasted R output did not receive full points; a proper discussion of the output, specifically the RMSE, was necessary, along with comparison of the RMSE to previous models. The best candidates were able to use a combination of both the RMSE and practical considerations to make a recommendation.

Though not necessary for full points, some candidates noted a weakness that applies to both ridge and LASSO is that the R implementation via glmnet restricts the models that can be used. In particular, the gamma/log model used earlier is not available. This may make this model more difficult to compare or compete with the previous model.

Many candidates made the correct observation that the glmnet implementation requires binarization of factor variables. For those who did not perform binarization in Task 6, this could be viewed as an advantage for this approach.

There are a variety of methods to reduce overfitting. I previously used an information criterion, BIC, to reduce model complexity. An alternative to reducing the number of variables used is to reduce the coefficients of each variable. This is done by adding a penalty to the loglikelihood that relates to the size of the coefficients. This diminishes the effect, particularly for features that have limited predictive power. There are two approaches to doing this, which come under the general term regularization. (A third approach, a combination of the two, will not be discussed here.) In both cases there is a hyperparameter to estimate that controls the extent of the reduction. This is normally selected through cross-validation. The specific methods explored are:

- Ridge regression: The penalty is proportional to the sum of the squares of the estimated coefficients. All of the coefficients are reduced but none are reduced to zero. Hence, all variables are retained.
- LASSO regression: The penalty is proportional to the sum of the absolute values of the estimated coefficients. All of the coefficients are reduced and some may be reduced to zero, effectively removing that variable.

Ridge regression is not recommended for this problem. Our goal is to identify variables that best predict Crash_Score and with all variables retained this approach will not be useful.

LASSO provides an alternative to forward and backward selection for variable selection. One advantage is that through cross-validation it selects the best hyperparameter using the same criterion (RMSE) that will ultimately be used to judge the model against unseen data.

Regularization methods requires binarization of categorical variables, so unlike the stepAIC performed earlier, which treated all factor levels of one variable as a single object to remove or retain in the model, the LASSO removes individual factor levels if they are not significant with respect to the base level.

In running the two regressions, ridge produced an RMSE of 6.2516 and LASSO produced an RMSE of 6.2463. The LASSO removed 20 factor levels. Note that 10 of them were related to the Month variable.

Based on the above considerations and the fact that neither regularization approach improved the RMSE, I recommend that the original regression be used.

Task 10 – Consider a decision tree (5 points)

Most candidates were able to list at least two advantages and two disadvantages of using decision trees. Candidates lost points for conflicting information in the two lists, for invalid or incomplete information, for poorly worded or poorly structured information, for forgetting that this was to be a comparison to GLMs and not just a fact list about trees, or for failing to relate these pros and cons to the business problem.

A regression decision tree is an alternative method of linking predictors to a target variable. A tree divides the feature space into a finite, non-overlapping set of buckets. All observations in a given bucket have the same predicted value. As with GLMs, there are methods to control overfitting such as cost-complexity pruning. The advantages of trees relative to GLMs are:

- They can be easier to interpret, provided there aren't too many buckets. As the name implies, a tree-like diagram can be constructed that indicates which observations get into the various buckets. Depending on the link function, the coefficients in a GLM may be difficult to explain and interpret.
- Categorical data is automatically handled. There is no need to binarize or determine a base class.

  The following additional advantages also received credit. This list is not exhaustive.

  - *Interactions are automatically handled. There is no need to identify potential interactions prior to fitting the tree.*
  - *Variables are automatically selected. Some variables simply do not appear in the tree.*
  - *They can produce non-linear relationships between the predictor variables and the response. (This is a weaker response because for this problem most all variables are categorical and hence linearity is not an issue.)*

Disadvantages relative to GLMs are:

- Even with pruning, there can be considerable overfitting to the training set.
- When underlying data changes, break points for decision trees can change significantly, leading to low user confidence in the model

  The following additional disadvantages also received credit. This list is not exhaustive.

  - *When fitting a single tree, the locally greedy algorithm is unlikely to find a globally optimal tree.*
  - *With continuous predictors, the bucketing of features means that some small changes can lead to a large change in the prediction while other small changes can lead to no change in the prediction. (This is a weaker response because for this problem most all variables are categorical and hence are already bucketed.)*

Task 11 – Executive summary (20 points)

The executive summary should concisely provide an overview of Tasks 1-8 and requires an explanation of the technical analysis using non-technical language appropriate for the client. It is acceptable for candidates to include technical terms as long as they are explained to the reader. The executive summary need not be written with a memo line at the beginning. However, it is necessary to identify the North Carolina Department of Transportation as the client. This can also be done in the body of the report. Many candidates fell short when describing the problem statement, the data, the model selection process, their interpretation of the model output, the business implications, and the next steps. Many candidates would have benefited from a clearer (or any) visual presentation of results.

A key to success is interpreting the results of Tasks 1-8 rather than summarizing every detail.

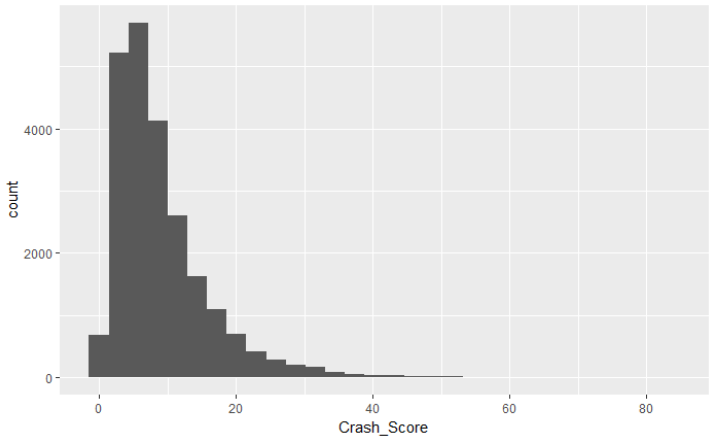
To: North Carolina Department of Transportation

From: Actuarial Consulting Firm

Subject: Exploration of factors relating to crash severity

This report provides a preliminary investigation into the factors that contribute to greater or lesser severity of automobile accidents. We have used the 2014-2019 Cary, NC data that you supplied. The severity of an individual accident is represented by Crash Score, a variable you have created that combines information from each accident such as number of injuries and number of vehicles involved. You also supplied variables that you believe relate to Crash Score. As requested, we have applied analytic techniques to determine which of those variables relate to Crash Score and the magnitude of the effect for those that do.

The data you supplied came from 23,127 accidents in the town of Cary, NC, from 2014 through mid-2019. The supplied data had no missing values or obvious errors. We have therefore assumed that it is reliable for the intended purpose. Because the data is from a single locality, the findings of this analysis may not generalize well to crash severity for the whole state. The distribution of Crash Score is presented below.



It is clear that Crash Score is skewed to the right, having mostly low values and a few very high values. This will be accounted for when constructing a model that predicts this variable.

> *The executive summary should provide a general description of the predictor variables and walk the audience through the high level steps taken to end up with the features used in the final model. For Task 1, it is sufficient to present the key results It should be noted that the differences in the target are fairly small.*

The supplied variables appear to be a good cross-section of prevailing conditions at the time of the accident. Some variables may prove relevant but are not within NC DOT's control. These are the weather, road conditions (e.g., dry or wet), and the year, month, and time of day. Other variables may be, to some extent, within NC DOT's control. They are road features (e.g., intersection or ramp), road character (e.g., straight or curved), road maintenance (e.g., state or U.S.), road configuration (e.g., one way or two way), road surface (e.g., concrete or asphalt), presence of street lamps, traffic control (e.g. signal or stop sign), and if the crash occurred in a work area.

A preliminary investigation showed that many of these variables appear to relate to Crash Score, but none have an obvious, strong effect causing it to stand out. In many cases, distinctions between various categories were slight both in meaning and effect, so these were combined in order to seek out broader distinctions. For instance, in Road Character, the distinctions among different types of straight roads seemed to have no effect on Crash Score, so these were grouped together as straight roads, and similarly for curved roads.

> *Because these variables were supplied by the client, they should know that some alterations were made. Failing to disclose these changes could lead the client to misinterpret the model results. The executive summary need not list every alternation made, but some indication of the nature of the alternations or an example should be provided.*
>
> *The final model in this solution did not use either the variable created by PCA nor the interaction term. It is not necessary to discuss explorations that do not affect the final model. However, the use of PCA removed two of the original variables from further consideration and that is worth communicating.*

Before beginning the formal analysis, one additional alteration was made to the data. Three of the variables, Road Conditions, Weather, and Light appear to be measuring similar quantities. An analysis of them indicated that the first two could be combined to create a new, numerical, variable that is strongly positive in wet and rainy conditions and strongly negative in dry and clear conditions. A new variable called WETorDRY was created to replace those two. The variable Light was left as is.

> *Many candidates failed to justify their model choice. The audience is unlikely to know about link functions and other components of a GLM. The point of this description is to motivate its use and provide confidence that a method was selected that is capable of solving the problem.*

With the data ready for analysis, we have selected a generalized linear model (GLM) for drawing inferences. A GLM has the advantages of being able to accommodate a variety of distributions for the quantity being predicted (hence can account for the skewness seen in the graph presented earlier), can ensure that predictions are always appropriate (positive numbers in this case), provides a systematic method for deciding which factors have predictive power, and, for those that do, provides an easily interpretable measure of the effect each factor has on Crash Score.

> *When describing the feature selection process the language used is not that of bias/variance, overfitting/underfitting, backward/forward, but rather a general view regarding why features should be removed and the motivation for the choices (in this case) of BIC and forward selection. Candidates who made choices in Task 6 based on business considerations had an easier time explaining them here.*

After selecting a GLM that aligns with the data, the next step is to use a statistical method to simplify the model by selecting only those factors that have predictive power. Commonly used statistical methods allow some leeway in making this selection. We have selected a method that tends to remove more variables. This will allow NC DOT to concentrate on the ones that have the greatest relationship to Crash Score. As a result, the following variables were retained.

- Road Configuration
- Road Feature

- Time of Day
- Traffic Control

The validation step is referenced, but no technical terms are used. The point here is to assure the reader that validation was done and it was successful.

Various tests, both numerical and graphical, were conducted to ensure that the model selected was appropriate. In particular, we verified that removing the other variables did not reduce predictive power and that the various assumptions underlying use of a GLM were reasonably satisfied.

Many candidates struggled to adequately explain the model output for the reader. The readers are not likely to be familiar with regression output, so copying R output here may not be helpful. Presenting a table with key items more effectively communicates what is important. The interpretation should be consistent with the model, in this case translating the coefficients into percentage changes. One alternative approach used successfully by some candidates was demonstrating how to use the model coefficients and intercept to calculate a Crash Score. The model results should be interpreted regarding how much sense they make.

When discussing the effect of a factor level, it is important to note that the change relates to the base class.

The GLM coefficients tell us the effect of various levels of the selected factors on Crash Score. Due to the nature of the model, the coefficients can be transformed to percentage changes.

Feature	Interpretation	Commentary
Road Feature = INTERSECTION-RAMP	70% increase in Crash Score compared to non-intersection-ramp	Accidents at intersections and ramps may be more likely to involve more than one car.
Time of Day = LATE-EARLY	4% decrease in Crash Score during 4am to 8am or 8pm to 12am compared to daytime (8am to 8pm)	During the 8pm to 8am period there may be less traffic, leading to more single-car crashes.
Time of Day = OVERNIGHT	11% decrease in Crash Score during 12am to 4am compared to daytime (8am to 8pm)	
Road Configuration = ONE-WAY	6% decrease in Crash Score compared to TWO-WAY-NO-MEDIAN.	One-way roads are unlikely to have head-on collisions and hence less severe crashes. It is not clear why the existence of medians leads to higher crash scores, perhaps cars go faster in this setting.
Road Configuration = TWO-WAY-PROTECTED-MEDIAN	6% increase in Crash Score compared to TWO-WAY-NO-MEDIAN.	
Road Configuration = TWO-WAY-UNPROTECTED-MEDIAN	44% increase in Crash Score compared to TWO-WAY-NO-MEDIAN..	
Road Configuration = UNKNOWN	2% increase in Crash Score compared to TWO-WAY-NO-MEDIAN.	

Traffic Control = CONTROLLED	7% increase in Crash Score with signal or stop sign versus no control.	While accidents may be less likely at controlled intersections, they are more unexpected, and thus involve higher speeds or more vehicles.

The concluding paragraph should point out that the effects are small, but that is up to NC DOT. Candidates produced many different, but acceptable, models. Conclusions may differ depending on the model created.

It is not appropriate to discuss alternative modeling approaches that might be tried given more time, such as random forest. The client expected you to build the best possible model with the available time and data.

Our investigation into crashes in Cary, NC has indicated factors that lead to changes in the expected severity. We note that the changes are not particularly large but may help guide NC DOT in making changes to road configurations. Should you find this report useful, we would be pleased to analyze a set of state-wide data.