Exam PA June 14, 2019 Project Statement

General information for candidates

This assignment has two components. One is a statement of the business problem to be addressed. The other is a list of tasks to be done. Your report will consist of responses to ten specific tasks followed by an executive summary. The audience for the task responses is the examination grading team. Hence, technical language can be used. Each task will be graded individually, so be sure any work that addresses a given task is done within the writeup for that task. The final item in your report is an executive summary written for an audience **not** familiar with analytics concepts.

This document and the report template indicate the points assigned to each of the eleven components. The total is 100 points. Each task will be graded on the quality of your thought process and conclusions. The executive summary will be graded on the quality of the presentation. Note that a component of the grading of the first ten tasks will also relate to the quality of the exposition, but these sections need not be written as formal reports.

At a minimum you must submit your completed report template and an Rmd file that supports your work. Graders expect that your Rmd code can be run from beginning to end. The code snippets provided should either be commented out or adapted for execution. Ensure that it is clear where in the code each of the tasks is addressed. Your thought process and conclusion for each task should be completely contained within your Word report. The Rmd code should be clear, contain commentary, and support your work.

You may submit other files as needed to support your work. In addition to Word (.docx) and RStudio (.Rmd) files, you may also submit Excel files (.xlsx or .csv). There is a limit of 10 files and no file can be larger than 25MB.

Business Problem

Your actuarial consulting firm has been hired by the North Carolina Department of Transportation to help them understand the factors that contribute to the severity of vehicle crashes. For this preliminary study they have obtained data from 2014-2019 on crashes in Cary, North Carolina (NC). If this investigation looks promising, they will provide statewide data for further analysis.

The dataset used in this assessment was provided by the town of Cary, NC, and is used with permission. Some cleaning was performed in advance. The data dictionary at the end of this document describes the available variables. The target variable *Crash_Score* combines several factors, such as the number of injuries and fatalities and the number of vehicles involved. There is no scale to this variable other than larger values indicate a more severe crash. There is no information about the frequency of crashes. It is not necessary for you to paste the dictionary into your report.

Your goal is to identify and interpret factors that relate to a higher or lower *Crash_Score*.

To get you started, your assistant has done some preliminary analyses, which are scattered throughout the supplied Rmd file. The analyses:

- Removed all entries with missing data (done prior to providing the dataset).
- Removed any outliers (done prior to providing the dataset).
- Releveled the factor variables so that the base level has the most observations.
- Ran a principal components analysis on selected variables.
- Split the dataset into training and testing sets.
- Constructed an ordinary least squares (OLS) regression model using the training set and all variables currently under consideration and measured its effectiveness by applying it to the test set and calculating the root mean square error (RMSE).
- Supplied code to perform various tasks as indicated in the Rmd file.

There is no assurance that your assistant has made the best choices in each code chunk. As an example, one of the code chunks for Task 2 produces means and medians. Perhaps other summary statistics would be more useful.

Specific Tasks

The tasks are intended to be done in order with results from one task informing work in later tasks. Graders will look for the solution to a given task within that task's area in the report and Rmd file.

In all cases you should justify the choices you make in your report.

When tasks 1-4 are complete, a set of features will have been identified for use in subsequent models. No additional features should be created after these tasks have been completed. This does not preclude removing features in the model-building process.

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

    Use graphical displays and summary statistics to form preliminary conclusions regarding which variables are likely to have significant predictive power.

2. (*5 points*) Reduce the number of factor levels where appropriate

    Several of the variables have a small number of observations at some of the factor levels. Consider using knowledge of the factor levels as well as evidence from Task 1 to combine some of them into factor levels with more observations.

    Do not reduce the number of levels for Rd_Conditions, Light, and Weather. These variables are addressed in the next Task. To ensure all candidates work with identical variables, they should not be changed in this Task.

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

    Your assistant has provided code to run a PCA on three variables. Run the code on these three variables. Interpret the output, including the loadings on significant principal components. Generate one new feature based on your observations (which may also involve dropping some current variables). Your assistant has provided some notes on using PCA on factor variables in the Rmd file.

4. (*7 points*) Select an interaction

Select one pair of features that should be included as an interaction variable in a generalized linear model (GLM). Do this by first proposing two variables that are likely to interact and then using the supplied boxplot function to confirm the existence of an interaction. Continue until a promising interaction has been identified. Do not use the features that were part of the PCA exploration in Task 3 when looking for interactions. Include your selected interaction when constructing a GLM in the following tasks.

Tasks 5-8 relate to constructing a GLM.

5. (*10 points*) Select a distribution and link function

Evaluate two potential combinations of distribution and link function for applying a GLM to the training dataset. (Typing ?family in the Console will provide help. Included are the combinations that can be used with the glm function.) Explain, prior to fitting the models, why your two choices are reasonable for this problem. Fit both models using the features developed in Tasks 1-4 and select the best combination, justifying your choice. Use only that model in Tasks 6-8.

6. (*12 points*) Select features using AIC or BIC

AIC and BIC are among the available techniques for feature selection. Briefly describe them and outline the differences in the two criteria. Make a recommendation as to which one should be used for this problem. Use only your recommended criterion when completing this task.

Some of the features may lack predictive power and lead to overfitting. Determine which features should be retained. Use the stepAIC function (from the MASS package) to make this determination. When using this function, there are two decisions to make. Make each decision based on the business problem. Use ?stepAIC to learn more about these parameters (note that the MASS package must be loaded before help on this function can be accessed).

- Use direction = “backward” or direction = “forward”
- Use AIC (k = 2) or BIC (k=log(nrow(train)))

7. (*6 points*) Validate the model

Run the model from Task 6 and evaluate the RMSE against the test set and compare it to the assistant’s OLS model. Also provide and interpret diagnostic plots to check the model assumptions.

8. (*9 points*) Interpret the model

Run the selected model from Task 6 on the full dataset and provide the output. Interpret the results in a manner that will provide useful information to the North Carolina Department of Transportation. This will be the model used in your executive summary.

9. (*12 points*) Investigate ridge and LASSO regressions

Code is provided to run both ridge and LASSO regressions. Use the features developed in Tasks 1-4. No other changes in parameters need be done. Compare the RMSE on the test set to that

from your model from Task 7. Note that the *glmnet* package is restricted in the model forms. The code provided predicts the target variable using a Gaussian distribution and the identity link. There is no need to try other combinations.

Provide an explanation of the difference in the three approaches (forward or backward as used in Task 6, ridge, LASSO). Which of the three would you recommend for this analysis? Do not base your recommendation solely on the mean squared errors from each model.

10. (*5 points*) Consider a decision tree

An alternative to the GLM is a regression decision tree. Do not create such a tree. Comment on the pros and cons of using a regression tree for this problem versus the GLM constructed in Task 6.

11. (*20 points*) Executive summary

Your executive summary should reflect the information provided and work from Tasks 1-8 as relevant to the North Carolina Department of Transportation. Your executive summary should include a problem statement and a coherent explanation of all the steps leading to your recommended model and conclusions.

Data Dictionary

Crash_Score	Measures the extent of the crash using factors such as number of injuries and fatalities, the number of vehicles involved, and other factors	A positive number with two decimal places
Year	Calendar year of the crash	Integer 2014-2019
Month	Calendar month of the crash	Integer 1-12 (1 = January, 12 = December)
Time_of_Day	Time of day, in four-hour blocks	Integer 1-6 (1 = midnight to 4am, 6 = 8pm to midnight)
Rd_Feature	Special feature of the road where the crash occurred	NONE – no special feature INTERSECTION – the meeting of at least two roads RAMP – exit or entrance ramp to a controlled access road DRIVEWAY – entrance to home or business OTHER
Rd_Character	Description of the road where the crash occurred	STRAIGHT-LEVEL – no curves or hills STRAIGHT-GRADE – no curves, but on a hill (up or down) STRAIGHT-OTHER CURVE-LEVEL – on a curve but no hill CURVE-GRADE – on a curve and on a hill CURVE-OTHER OTHER

Rd_Class	Classification of the road type	STATE HWY – Maintained by the state government US HWY – Maintained by the federal government OTHER
Rd_Configuration	Design of the road	TWO-WAY-PROTECTED-MEDIAN – Traffic in both directions, separated with a barrier TWO-WAY-UNPROTECTED-MEDIAN – separated but with no barrier TWO-WAY-NO-MEDIAN – no separation ONE-WAY UNKNOWN
Rd_Surface	Material used for the road surface	SMOOTH ASPHALT COARSE ASPHALT CONCRETE GROOVED CONCRETE OTHER
Rd_Conditions	Condition of the road	DRY WET ICE-SNOW-SLUSH OTHER
Light	Lighting	DAYLIGHT DARK-NOT-LIT – no street lamps in area DARK-LIT DUSK DAWN OTHER
Weather	Weather conditions	CLEAR RAIN CLOUDY SNOW OTHER
Traffic_Control	Any items that control traffic flow	SIGNAL – lighted stop/go signal STOP-SIGN YIELD NONE OTHER
Work_Area	Was the crash in a work area?	YES/NO