Sample Project- Hospital Readmissions

Note to candidates concerning this sample project – This sample project takes a different approach from the one used in the other sample problems and in the December 2018 exam. The approach used here matches what you can expect to see in the exams given in June 2019. For the actual exams, the number of tasks and point values may differ. Also, the techniques being requested may differ.

The sample solution represents a high-quality solution. Note that there are multiple decision points and alternative decisions may be equally acceptable. Graders look for an understanding of the issues that relate to the decision and how they are applied to the business problem.

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

Predictive models for hospital readmission rates are in high demand due to the U.S. Centers for Medicare and Medicaid Services (CMS) Hospital Readmission Reduction Program.[1] In 2010 CMS

[1] Medicare is a health care program provided by the U.S. government. It is primarily for people age 65 and older. However, benefits are available for younger people under certain circumstances.

introduced a penalty for excess readmissions: when admissions are excessive, they charge a percentage of a Hospital's *total* Medicare revenue, not just the revenue for readmissions that are excessive. Thus, the penalty can be a significant adjustment to the hospital's bottom line.

It is therefore desirable to be able to identify patients who are most likely to be readmitted. Doing so would allow the hospital to take proactive actions with these patients to reduce their likelihood of being readmitted.

The current approach is to use the LACE index for predicting which patients are at risk of readmission. LACE is one of the most popular predictive tools among hospitals in the United States. The LACE index is a simple tool with four parameters: **L**ength of stay, **A**cuity of admission, **C**omorbidity, and **E**mergency visits in the previous 6 months. This project entails building a model using patient level data with more specific patient information. The AUC Value for the LACE index is reported at 0.70.

You have been hired by a group of hospitals to use the data they have supplied to obtain a generalized linear model (GLM) that is superior to the LACE index. This will be demonstrated by achieving an AUC value greater than 0.70. There is also interest in an alternative validation measure that will be described later. The performance of the LACE index on this validation measure is not known.

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

- Removed all entries with missing data (done prior to providing the dataset)
- Provided code chunks that can be used for
  - Binarization of factor variables
  - Splitting the dataset into training and testing sets
  - Combining factor levels
- Releveled the factor variables so that the base level has the most observations
- Provided code to run a *K*-means cluster analysis on selected variables
- 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.

For this assignment, do not perform undersampling or oversampling.

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. (*6 points*) Perform univariate exploration of the four non-factor variables

   Use graphical displays and summary statistics to determine if any of these variables should be transformed and, if so, what transformation should be made. Do your recommended transformations, if any, and delete the original variables.

2. (*5 points*) Examine relationships between DRG.Class and DRG.Complication

   It appears there is some overlap in these two variables. Examine them and determine if it makes sense to combine them in some manner. If so, create the new feature and delete any variables that will no longer be used.

3. (*9 points*) Use observations from cluster analysis to consider a new feature

   Your assistant has provided code to run a *K*-means cluster analysis on LOS and Age. (If you elected to transform either of these variables in Task 1, use the transformed version.) The analysis sets nstart = 1, something your assistant was confused about. Explain the role of this parameter and, if appropriate, change the value.

   Run the code on these two variables. Select the best number of clusters. Create a factor variable that could be used in place of these two variables, as will be explored in Task 6. Your assistant has provided code that illustrates how to do this.

4. (*5 points*) Select an interaction

   Select one pair of features, from among Gender, Race, ER, and HCC.Riskscore, that should be included as an interaction variable in your GLM. Do this by first proposing two variables that are likely to interact and then using one of the supplied functions to graphically confirm the existence of an interaction. Continue until a promising interaction has been identified. Include your selected interaction when constructing a GLM in the following tasks.

Tasks 5-8 relate to constructing a GLM.

5. (*8 points*) Select a link function

   With the target variable being only 0 or 1, the binomial distribution is the only reasonable choice. Your assistant has done some research and learned that for the glm package in R there are five link functions that can be used with the binomial distribution. They are shown below (the inverse of the link function is presented here as it represents how the linear predictor is transformed into the actual response), where η is the linear predictor and *p* is the response.

   - Logit (link = "logit") $p = \frac{e^{\eta}}{1+e^{\eta}}$
   - Probit (link = "probit"): $p = \Phi(\eta)$ where Φ is the standard normal cumulative distribution function
   - Cauchit (link = "cauchit"): $p = \frac{1}{\pi}\arctan(\eta) + \frac{1}{2}$ (this is the cumulative distribution function of the standard Cauchy distribution, which is a *t* distribution with one degree of freedom)

- Log (link = “log”): $p = e^{\eta}$
- Complementary log-log (link = “cloglog”): $p = 1 - e^{-e^{\eta}}$

Identify the best choice of link function. Explain, prior to fitting the model, why your choice is reasonable for this problem. Fit your choice plus one alternative and compare your results. You can switch to the alternative if it performs better.

With respect to Task 3, use the two original LOS and Age variables (or their transformed versions) and do not use the feature created from the cluster analysis. Do use any transformations from Task 1, any replacement variable created in Task 2, and the interaction from Task 4.

6. (*5 points*) Decide on the factor variable from Task 3

Fit an additional model retaining everything from Task 5, except drop LOS and Age (or their transformed versions) and add the variable created via clustering. Select the best model from these two, justifying your choice.

Use the model you select here in Task 7.

7. (*15 points*) Select features

The remaining model construction task is to remove variables to prevent overfitting. One approach is to use a function such as stepAIC on the current set of variables. However, for factor variables this function either retains the variable with its existing levels or removes the variable entirely. It does not allow for the possibility that individual factor levels may be insignificant with regard to the base level (and hence could be combined with it) or insignificantly different from other level(s) (in which case they could be combined into a new level).

Simplify the model by removing features entirely or combining factor levels as appropriate. Use an approach that relies on hypothesis tests or information criteria (such as AIC). Do not use a regularization method such as LASSO. Be sure and explain your methodology and why it is a reasonable approach.

Calculate the AUC against the test set and compare it to the LACE index.

When finished with this task, this will be your final model form.

8. (*6 points*) Interpret the model

Run the selected model from Task 7 on the full dataset and provide the output. Interpret the results in a manner that will provide useful information to the hospitals. This will be the model discussed in your executive summary.

9. (*9 points*) Set the cutoff

The hospital is also interested in an alternative approach to understanding the benefits of using your model. They have produced the following approach to evaluating the benefits.

- Any patient who is readmitted incurs a cost of 25 to the hospital (these numbers have been scaled).
- An intervention program is being considered such that for a cost of 2 any discharged patient who receives the intervention will not be readmitted. The hospitals are considering three options:
  - Do not employ the intervention program.
  - Apply the intervention to every discharged patient.
  - Apply the intervention only to those patients predicted by the GLM to be readmitted.
- The hospital does not currently have information about how the LACE index would perform if used to decide which patients receive an intervention.

Use the confusion matrix from running the model on the full dataset to estimate the cost under each of the three options. You should change the cutoff for predicting readmission to optimize this result.

10. (*12 points*) Consider alternative models and model construction techniques

There are alternatives to the approaches used above. For each one presented below, indicate the advantages and disadvantages of using it for this business problem.

- LASSO regularized regression
- Classification tree using cost-complexity pruning
- Random forest

11. (*20 points*) Executive summary

Your executive summary should reflect the information provided and your work from Tasks 1-9 as relevant to the hospitals. 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

Readmission.Status	The target variable, it is 1 for patients that were readmitted, 0 otherwise
Gender	M indicates male, F indicates female
Race	There are four categories: Black, Hispanic, Others, White
ER	The number of emergency room visits prior to the hospital stay associated with the readmission, an integer
DRG.Class	Diagnostic Related Group classification. There are three categories: MED (for medical), SURG (for surgical), UNGROUP.
LOS	Length of hospital stay in days, an integer
Age	The patient's age in years, an integer. (Note that while most Medicare recipients are age 65 or older there are circumstances in which those under 65 can receive benefits.)
HCC.Riskscore	Hierarchical Condition Category risk score. It is designed to be an estimate of a patient's condition and prospective costs. It is a continuous variable, rounded to three decimal places. Higher numbers indicate greater risk.
DRG.Complication	Complications, with five levels: MedicalMCC.CC, MedicalNoC, Other, SurgMCC.CC, SurgNoC. MCC.CC complications or comorbidities that may may be major. NoC means no complications or comorbidities.