MXB334: Assessment 3 Description

|  |  |
| --- | --- |
| Task | Credit Risk Model |
| Unit Learning Outcomes Addressed | 1. Expertly and critically carry out statistical analysis using statistical models in the analysis of various data sets and examples. 2. Use R to carry out statistical analyses. 3. Communicate statistical conclusions clearly and concisely both in written form and orally. |
| Due date | Week 14/15 |
| Weighting | 40% |
| Specifications | Groups of 3 or 4 |
| Overview | |
| This project will assess your ability to work as a team to apply generalised linear modelling to create a reproducible piece of analysis of real-world customer loan data. You have 2 opportunities to discuss the project with real analysts working for the Bank of Queensland. The analysis will culminate in the proposal of a credit risk model and an evaluation of its performance using industry standard measures and techniques.  The way you will work will closely mirror that of real analysts working in the financial sector. You will be expected to plan your project, give status updates, summarise your findings for executives, document your analysis in a reproducible way, use version control/collaboration tools, and review your performance against you measures of success.  The project is comprised of deliverables that will be due at various stages throughout semester. They are:   1. A Delivery Plan that decomposes the project into a number of tasks or milestones spread over the semester, to be reviewed with your lecturer weekly. 2. An Interim Milestone Report that details how the dataset for modelling has been created from the initial data sources. 3. A Summary on a Page (SOAP) distilling the key outcomes of your project down to a 1 page executive summary. 4. A final Reproducible Analytical Report that formulates, validates, and tests your proposed credit risk model. 5. A Performance Review where you will argue for the grade your work should receive based on the assessment criteria. | |
| What you will do | |
| The overarching task is the creation of a statistical model for loan default. You will use binomial regression within the generalised linear modelling framework to build this model. Models like the one you will build are used by money lending organisations to assign probability of default to loan applications. This information allows effective management of a lending (credit risk) portfolio.  You will plan your work from the project outset and update the plan as you complete tasks or come to better understanding of the remaining work.  Initially you will work toward the interim milestone, which will be to complete preparation of the data for analysis. This will involve:   * joining, filtering, and cleaning data form 2 sources * documenting your decisions on how to handle missing and unusual data * defining and creating the response variable * excluding variables contaminated by information relating to the response * applying variable selection algorithms or heuristics taught in the course to narrow down the large number of covariates for modelling. * Combining all of the above into a reproducible interim report written in Rmarkdown.   Using your dataset, you will build a model for risk of loan default using binomial regression. This will involve:   * Exploratory analysis of variables of usual interest in credit risk.   + It will be up to you to find out what these are from Bank of Queensland analysts. * Selecting from a space of candidate models using goodness of fit measures and cross validation (AIC, Gini/AUC). * Selecting among 3 link functions: logit, probit, cloglog   + With justification for final selection * Testing and validating the assumptions of the model * Evaluating the performance of the model against existing benchmark * Drawing conclusions about the effects of important variables   + Makes statements of the uncertainty relating to the effects * Combining all of the above into a reproducible report written in Rmarkdown.   Using the insights gained from your analysis you will distil the interesting results into a Summary On a Page (SOAP) for executive consumption. You will:   * Use visuals to communicate where appropriate * Highlight Important covariates in the model * State key assumptions   + in particular any assumptions that could not be verified. Highlight these as risks. * State state performance of model   + Compare it to benchmark in terms that relate directly to the executive.   For the final part of this assessment you will submit an individual performance review where you will assess the project and yourself against performance criteria. A template form shall be provided to you on Blackboard. The argument you make for the grade you shall receive will be taken into consideration in the grade awarded to you. | |
| What you will submit | |
| The **Delivery Plan** is a living document that maps the known remaining tasks in this project onto weeks in the project timeline. It is a living document that you will update weekly. Tasks may be added or removed at any time. Each week during the Thursday practical session, your group will make a short report on how you are tracking toward planned items with your lecturer. This report is an opportunity to raise road blocks the group has encountered and get help.  The **Interim Milestone Report** and the **Reproducible Analytical Report** are collaborative works that will be housed in a private Github repository that will be issued to your group. You will use Github’s workflow to collaborate and manage shared authorship in Rmarkdown format. At the respective submission deadlines an image will be made of your group’s repository for grading. Any instructions required to reproduce your report should be entered into the repository’s readme file, although this should not be necessary.  For the **Interim Milestone Report**, the repository will be imaged on **5pm Friday of Week 9**.  For the **Reproducible Analytical Report** the repository will be imaged on **5pm Friday of Week 14**.  You are encouraged to use Github and Rmarkdown for your **SOAP**, however you may use any format. Showcase your design skills if you wish! There will be a submission link placed on Blackboard to use if you do not wish to included it in your Github repository.  The deadline for submission (or push to Github) of the **SOAP** is **5pm Friday of Week 14**.  Your **Performance Review** needs to be submitted individually on Blackboard via an assignment submission link.  The deadline for submission of the **Performance Review** is **5pm Friday of Week 14.** | |
| Resources and Useful References | |
| 1. Logistic Regression Lecture Notes 2. Cross validated and Stack Overflow websites. 3. Rmarkdown documentation: <http://rmarkdown.rstudio.com/> 4. R data wrangling cheatsheet: <https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf> 5. Github workflow guide: <https://guides.github.com/introduction/flow/> 6. Blackboard Folder for this Project. | |

# A Top Level Effort

## Delivery Plan

* Updated weekly with new/revised tasks and realistic allocations work to weeks.
* Contains all tasks to be completed.

## SOAP

* Clear evidence in design of visualisations with intent to communicate information relevant to analysis.
  + Axis labels, axis scales, legends that support this.
* Clear explanation or visualisation of the uncertainty around analysis outputs that are relevant to the executive.
* May name methods used but avoids inappropriate explanations of them.
* Unambiguously states the outcome of the analysis in a way that directly relates to Executive.
* Refers to performance relative to existing benchmark using terms relevant to Executive, e.g. x% more loans or more $ with y% less defaults..
* Relative importance of covariates clearly communicated.
* Calls out any risks.
* Clearly states key assumptions.
* Contains link or reference to the repository of the analysis report.

## Interim Milestone Report

* Uses join functions to combine data sets based on key values
* Detects unusual observations using exploratory analysis
* Uses automated tools to identify and exclude covariates with low variance and high correlation.
* Applies critical reasoning to identify and remove covariates with that are contaminated by response.
* Justifies removal of:
  + categorical variables applying to small numbers of observations
  + variables with high proportion of missing/NA

## Reproducible Analytical Report

* Exploratory analysis creates intuition for results of statistical modelling.
* Makes justification for assumptions in terms of model form.
  + Link function choice backed up by tests.
  + Uses AIC or Gini to discriminate between nested models
* Clearly assesses quality of statistical model fit and validity of assumptions
  + Pearson Residual Chi-Sq test or Test of residual deviance
  + Leverage plots
  + Highlights issues and makes sound arguments for their resolution either backed up by theory or references.
  + Uses Cross validation with appropriately sized training, test, and validation sets.
* Documents model performance with ROC curve Confusion matrix, AUC, Gini.
  + Explains what the expected impact of model is if adopted, using validation set as justification.
* Clearly states effects (with uncertainty) of covariates:
  + Important in the chosen model
  + Of usual interest in credit modelling
* Has all required R code embedded in Rmarkdown document in relevant places, but does not make inordinate amounts of code visible in final output.
* Is inherently reproducible.
  + Can be executed and knitted from an empty environment without errors or numerous warnings.

## Performance Review

* Responds to template with concrete examples of contributions and measures against relevant benchmarks.

# A Passing Effort:

## Delivery Plan

* Updated most weeks.
* Some tasks are unrealistic for the allocated time frame.
* Some tasks are missing.

## SOAP

* Visualisations communicate information relevant to analysis but may be presented in a way not making this immediately obvious.
* Uncertainty communicated but not addressed directly, i.e. in a table or plot without proper explanation.
* Contains some mention of methods and concepts irrelevant to executive.
* States the outcome of the analysis in a way that leaves the Executive to do some translation to his or her own terms.
* Refers to performance relative to existing benchmark in a way that leaves the Executive to do some translation to his or her own terms.
* Importance of covariates identified but relative importance unclear.
* Ignores some risks.
* Clearly states key assumptions.
* Contains link or reference to the repository of the analysis report.

## Interim Milestone Report

* Uses join functions to combine data sets based on key values
* Misses some unusual observations using exploratory analysis
* Uses automated tools to identify and exclude covariates with low variance and high correlation.
* Applies critical reasoning to identify and remove some of the covariates with that are contaminated by response.
* Removes by fails to justify:
  + some categorical variables applying to small numbers of observatioSns
  + variables with high proportion of missing/NA

## Reproducible Analytical Report

* Exploratory analysis undertaken but lacks commentary that sets up expectations for the modelling.
* Makes justification for assumptions in terms of model form.
  + Link function choice backed up by tests.
  + Uses AIC to discriminate between nested models
* Clearly assesses quality of statistical model fit and validity of assumptions
  + Pearson Residual Chi-Sq test or Test of residual deviance
  + Leverage plots
  + Highlights issues and makes sound arguments for their resolution either backed up by theory or references.
  + Uses Cross validation with training, test, and validation sets.
* Plots model performance with ROC curve AUC, Gini without commentary.
* States effects (with uncertainty) of covariates in a way that may require some translation. Considers:
  + Important in the chosen model
  + Of usual interest in credit modelling
* Has all required R code embedded in Rmarkdown, too much code visible in final output.
* Is inherently reproducible.
  + Can be executed and knitted from an empty environment without errors or numerous warnings.

## Performance Review

* Responds to template but lacks concrete examples of contributions.