ACST4062/8032

Actuarial Data Analysis

Assignment Two

This assignment is worth 20% of your final grade.

This assignment is out of 50 marks.

Assignments are to be submitted using Turnitin on the course Wattle site. Submitted assignments must include the cover sheet provided in this document. Please keep a copy of the assignment for your records.

The ANU is using Turnitin to enhance student citation and referencing techniques, and to assess assignment submissions as a component of the University's approach to managing Academic Integrity. University policies on plagiarism will be strictly enforced.

Background

You are an actuary working in the analytic team of Providence Bank, a major bank in a developed country.

During the coronavirus (COVID-19) pandemic, many individuals in the country may find themselves in financial distress, i.e., unable to make payments on loans and credit cards. Providence Bank is now seeking ways to support its loan customers who are affected by the pandemic and its economic impacts. Providence would like to build a predictive model to identify the loan customers who are most likely to be experiencing financial distress.

Providence recently finished a data collection campaign to collect data tracking the financial distress status of a sample of customers in Mulberry City, a major city currently in lockdown due to a spike in community transmissions of coronavirus. It plans to use this data to launch a targeted campaign to identify the most at risk loan customers and discuss options available for them. This is because by identifying and offering suitable options for these customers, Providence could lower the expected loss from them compared to the scenario of not identifying them and letting them default. However, Providence would also like to minimise contacting loan customers who are not in financial distress – to avoid distributing unnecessary benefits (which would lower Providence's profitability).

Questions

The *first_campaign.rds* dataset is the processed data from the campaign that contains the financial distress status of 5,000 loan customers who reside in Mulberry City (see data description for more information). You have been asked to build a predictive model based on this data to identify the loan customers who are most likely to be experiencing financial distress in Mulberry City, and prepare an accompanying report which details the modelling process and analysis. The report is directed to your team leader, Michelle.

Following a meeting with Michelle, you gathered that she would like you to conduct an exploratory analysis, and investigate the performance across various linear models and tree-based models (see the report outline for more details) before providing a final recommended model. Besides, Michelle requested a section in the report discussing the variables that contribute to the 'predictive power' of the models.

Based on her prior meeting with the financial distress specialist team, Michelle believes the table below could be a good starting point of capturing the marginal costs of making contact or no contact for all customers.

		Customer's status	
		Not distress	Distress
Action	Not contact	0	3
	Contact	1	0

The *second_campaign.rds* dataset is one of the datasets that would be used to identify the customers to be contacted. It contains information of another 3,000 loan customers who reside in

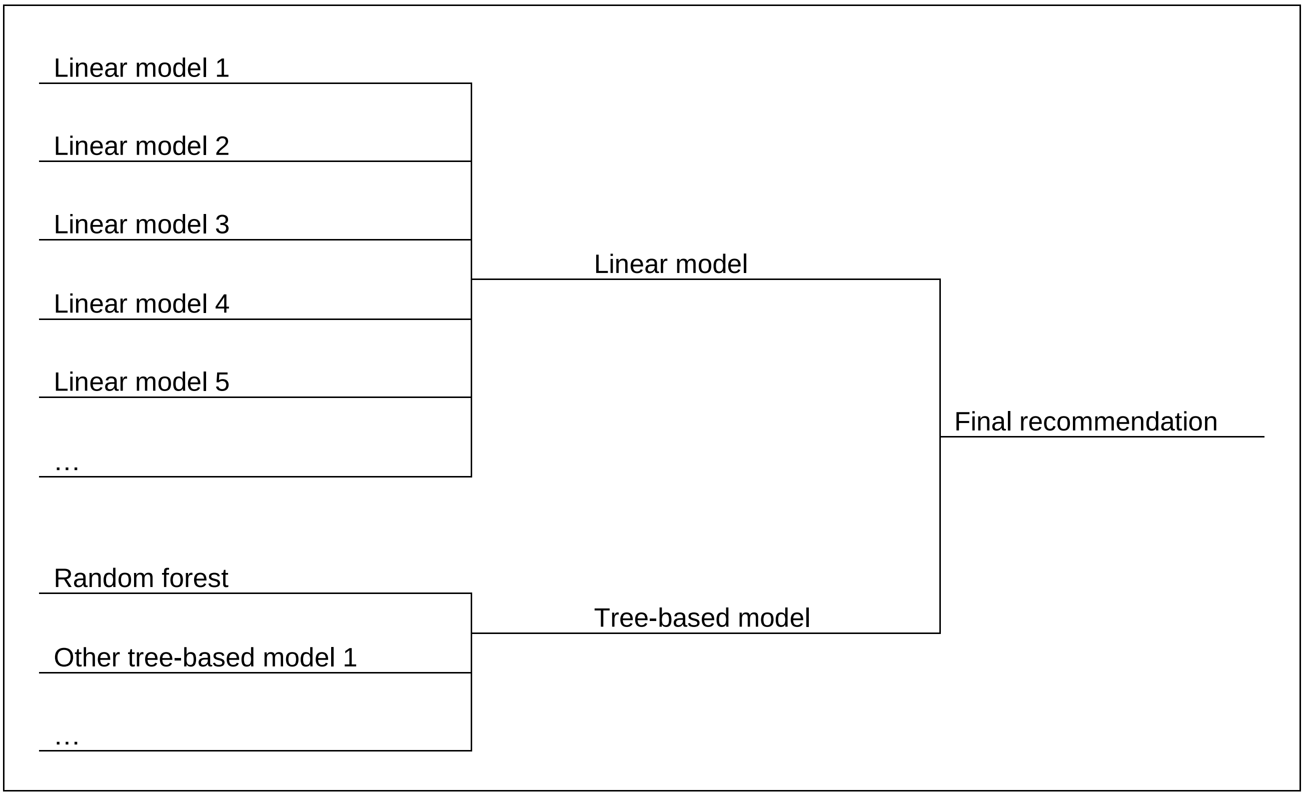
Mulberry City. Michelle would like you to discuss the prediction process of obtaining a list of customers who should be contacted, addressing the following queries:

1. How many customers should we be contacting?
2. What is the proportion of customers in financial distress among them? Note, “them” here refers to the customers whom the model suggests we should contact.

Here’s the proposed outline of the report:

- Exploratory analysis (5 marks)
  - Use a combination of words, graphs and/or tables to explain your findings.
- Linear models (15 marks)
  - Compare at least five distinct linear models, and from which determine a best linear model for the application. Explain your analysis and justify any modelling choices (e.g., models and tuning parameters to consider, resampling technique, performance metrics, etc).
  - Note, Ridge regressions with five distinct lambda values are counted as five distinct linear models. Without proper justification, this could appear to be a poor selection of models for comparison as it does not consider linear models with lasso penalty.
- Tree-based models (15 marks)
  - Compare a random forest model with at least one other tree-based model, and from which determine a best tree-based method for the application. Explain your analysis and justify any modelling choices (e.g., models and tuning parameters to consider, resampling technique, performance metrics, etc).
  - Note, a tree-based model can be a decision tree, a random forest (with different number of mtry), bagging (random forest with mtry equals the number of predictors), and a gradient boosting tree method.
- Final model and predictions (15 marks)
  - Discuss the predictive ability of the variables.
  - Select a final model and describe how this model can be used for prediction
  - Address Michelle’s queries regarding the customers in *second_campaign.rds* dataset.

The modelling flowchart Michelle and you agreed upon is given below:



Data descriptions:

The dataset consists of data collected from two sources – internal data and external credit report data. The internal data are the information collected during the loan application, except loan_status, which is collected during the first data collection campaign. The external credit report data are the variables are obtained from the customers' credit reports at the time of loan application. The credit report is provided by a credit bureau. Credit bureaus are the bodies which hold information on both consumer and business credit history. They supply credit providers with this information when the credit provider makes an inquiry as part of a credit check when a customer applies for credit (e.g. credit card or loan). Most numeric data have been grouped to protect the privacy of the customers. Except for income_joint and debt_to_income_joint, all data refers to the information about the primary customers.

Variable name	Description
Internal data	
loan_status	Loan status. This is the data collected from the first data collection campaign. 0 = "Not distress", 1 = "Distress".
loan_amount	Loan amount
loan_purpose	Loan purpose
income	Income of the customer. For joint application, it is the income of the primary customer.
debt_to_income	Debt-to-Income ratio. For joint application, it is the income of the primary customer. It is a number that represents the customer's total monthly debt obligation divided by his/her total monthly income. For example: If your monthly debt payments are $1,000 and your gross monthly income is $5,000, your debt-to-income ratio is 20%.
emp_length	Employment length 1 = "< 2 years", 2 = "10 + years", 3 = "2 - 5 years", 4 = "6 - 9 years", 5 = "unemployed"
application_type	Application type – single or joint loan application
income_joint	The combined income of the customers
debt_to_income_joint	The combined debt-to-income ratio of the customers.
home_ownership	Home ownership status
region	Region of residence in Mulberry City
External credit report data	
credit_score	It is a number that is calculated by credit scoring methods using credit information reported about the customer by credit providers (e.g. lenders, banks and other financial institution). A customer with high credit score is expected to be less likely to incur negative credit events (e.g. late payment or default) than a customer with low credit score.
credit_accounts	Number of current active credit accounts (e.g. credit card, loan, home mortgage, car loan or other credit).
recent_inquiry	If there is any inquiry in the past 12 months. An inquiry is a check into the customer's credit report by a company or individual. It typically occurs when the customer has applied for a credit account. How inquiries can

	affect a credit score varies depending on the frequency and recentness of enquiries, the type of credit applied for, and the type of credit provider.
delinquent	When was the last delinquent? A delinquent occurs when the customer falls behind on making required monthly payment on credit accounts (e.g. loans or credit card). Being late by more than a month is considered delinquent. The delinquent status about the customer is reported to the credit bureaus by the customer's credit providers.
credit_utilization	It is a ratio of the amount of revolving credit the customer is currently using divided by the total amount of revolving credit he/she has available. Credit utilisation ratio is based solely on revolving credit (e.g. credit cards and lines of credit). It does not include installment loans like mortgage or car loans. For example, suppose you have a credit card and a line of credit, each with a limit of $5,000, for a total credit limit of $10,000. Let's say you owe $1,000 on the line of credit and $2,000 on the credit card, for a total of $3,000 owed. Given this, your credit utilization ratio is 30%.
past_bankrupt	Whether or not the customer has had a bankruptcy.

Further instructions:

Cover page: Please include a cover sheet in your submission. The cover sheet is available on Wattle and it does not count towards the page limit.

Page limit: Please submit your assignment in a word or PDF document not more than 15 pages. You should think about how to best display your answers, workings and assumptions within this limit and marks will be awarded for presentation and communication. You may (and should!) include relevant tables and graphs in your document. Please ensure you explain your analysis and justify any modelling choices.

Appendix: Please submit the relevant R code used for modelling in the appendix. The appendix does not count towards the page limit. The appendix will not be marked but might be checked to clarify a response in the report if necessary.

Reference: You do not need to reference any other material to complete this assignment but if you do, please ensure you properly reference your work. You must adhere to appropriate practices regarding referencing the work of others in any work that you do: Accepted academic practice for referencing sources that you use in presentations and assignments can be found via the links on the Wattle site.

End of assignment

Note: All entities depicted in this assignment are entirely fictitious.