STAT 528 midterm exam

your name

Due at 11:59 PM on 3/14/2025

Analyze the SBA loans dataset

This dataset contains 899164 observations and 27 columns. This is historical data about actual business loans covered by the Small Business Administration (SBA) primarily from years 1970-2013 with emphasis on whether those businesses defaulted (charged off) or not (paid in full) on those loans. The observations are small businesses that seek loans to fund their operations, start-up costs, materials, payroll, rent, etc. The SBA works with banks by guaranteeing a portion of the loan to relieve banks of assuming all financial risk. To load the dataset in R, use the following command:

```
library(data.table)
sba = fread("https://uofi.box.com/shared/static/vi37omgitiaa2yyplrom779qvwk1g14x.csv",
                      header = TRUE, stringsAsFactors = FALSE)
```

Your assignment is to analyze this SBA dataset with the goal of investigating important variables for modeling the proportion that a loan is paid in full. Your analysis needs to be well-motivated. Any final model reported needs to be justified, properly validated, and well-fitting. Your final model should exhibit good classification of loans. You need to check modeling assumptions for any final models that you report using quantitative measures and visual diagnostics. You are allowed to consider subsets of the data as long as subsetting is well motivated. You are allowed to transform variables and create new variables provided that these manipulations are well motivated. Your analysis should be multifaceted, interesting relationships between variables should be reported. You are allowed to use materials from outside this course provided that you have a good reason for doing so and have considered the materials in this course (for example, if you consider flexible machine learning methods then you need to consider interaction terms in the glms). You are encouraged to add outside variables that may be important (economic measures for example, there have been several economic downturns over the range of data collection).

Note on selection bias: It is believed that the inclusion of loans with disbursement dates after 2010 would provide greater weight to those loans that are charged off versus paid in full. More specifically, loans that are charged off will do so prior to the maturity date of the loan, while loans that will likely be paid in full will do so at the maturity date of loan (which would extend beyond the dataset ending in 2014). Since this dataset has been restricted to loans for which the outcome is known, there is a greater chance that those loans charged off prior to maturity date will be included in the dataset, while those that might be paid in full have been excluded. It is important to keep in mind that any time restriction on the loans included in the data analyses could introduce selection bias, particularly toward the end of time period. This may impact the performance of any predictive models based on these data.

Note on ChatGPT and other large language models: These tools are allowed. However, anything that you write down which is not supported by your analysis, is directly contradicted by your analysis, is in conflict with the data, or is a logical contradiction (for example, reporting that a main-effect model is too simplistic to be realistic even though you are fully capable of including interaction terms) will **receive a 10 point deduction**. These deductions will apply even if you do not use ChatGPT or other large language models. While such errors often result from AI-generated responses, they can also stem from human oversight.

The original source for this data is "Should this loan be approved or denied?: A Large dataset with class assignment guidelines" by Min Li, Amy Mickel, and Stanley Taylor (https://www.tandfonline.com/doi/full/1

0.1080/10691898.2018.1434342). You are not allowed to directly copy the analyses in this reference. However, you can consider variables or transformations of variables that they considered. A description of key variables in this dataset are included on pages 3 and 4.

You should save your midterm as **netid_midterm** and it should be stored in a directory titled **midterm**. Do not include the dataset in your submission. Points will be deducted if your report is difficult to read or is too long. For example, a 50+ page document is seldom necessary and instead instills the impression that you are not confident in your report and are trying to overwhelm the reader or are simply being lazy. Keep in mind this famous quote:

> "If I had more time, I would have written a shorter letter.[1]"

That said, it is completely acceptable and encouraged to acknowledge shortcomings in your analysis and to discuss what you would have considered with more time. This can include a call for additional domain knowledge in addition to further statistical analyses. The primary focus should be on your final models and their justification. While you may briefly describe models that were discarded, detailed documentation of the model-comparison process is unnecessary and may be distracting. A short paragraph summarizing the rationale for rejecting alternative models is sufficient. However, exploratory analyses about how variables relate to the response are encouraged. You are not required to explore materials beyond the scope of this course, but you are encouraged to do so if you choose, especially if they provide meaningful comparisons to or complement your final model.

This midterm is worth **150 points**.

[1] Note the irony of this exam prompt being longer than necessary due to a wordy call for conciseness.

Here is a description of the variables:

Variable name	Data type	Description of variable
LoanNr_ChkDgt	Text	Identifier–Primary key
Name	Text	Borrower name
City	Text	Borrower city
State	Text	Borrower state
Zip	Text	Borrower zip code
Bank	Text	Bank name
BankState	Text	Bank state
NAICS	Text	North American industry classification system code
ApprovalDate	Date/Time	Date SBA commitment issued
ApprovalFY	Text	Fiscal year of commitment
Term	Number	Loan term in months
NoEmp	Number	Number of business employees
NewExist	Text	1 = Existing business, 2 = New business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise code, (00000 or 00001) = Nofranchise
UrbanRural	Text	1 = Urban, 2 = rural, 0 = undefined
RevLineCr	Text	Revolving line of credit: Y = Yes, N = No
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Currency	Amount disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan status charged off = DCHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approvedloan

(You may want to use regular expressions and pattern replacement functions such as `gsub` to convert currency variables to numeric.)

Here is a description of the first two digits of the NAICS classifications:

Sector	Description
11	Agriculture, forestry,fishing and hunting
21	Mining, quarrying, and oil and gas extraction
22	Utilities
23	Construction
31–33	Manufacturing
42	Wholesale trade
44–45	Retail trade
48–49	Transportation and warehousing
51	Information
52	Finance and insurance
53	Real estate and rental and leasing
54	Professional, scientific, and technical services
55	Management of companies and enterprises
56	Administrative and support and waste management and remediation services
61	Educational services
62	Health care and social assistance
71	Arts, entertainment, and recreation
72	Accommodation and food services
81	Other services (except public administration)
92	Public administration