GEORGE MASON UNIVERSITY | School of Business

MIS 431: Data Mining for Business Applications

Spring 2019

Final Project Instructions

The final project will be a chance for you to perform an applied data mining analysis on a real data set. We will be working with the **loan_data** data frame in this project. You can work on this assignment individually or as a group with another student in the class (no more than 2 students per group is allowed).

This dataset contains information on over 3,000 individuals who secured a personal loan from a national lender in the United States. The description of this data and the variables contained in it are provided below.

The final project is worth **100 points** and has the following four key deliverables:

1. (**10 Points**) The **R notebook file** that produces the data mining results and analysis
2. (**40 Points**) Exploratory Data Analysis (EDA) and Variable Importance Analysis
3. (**40 Points**) Predictive Modeling – Three Methods of Classification
4. (**10 Points**) Summary of Results and Recommendations to Decision Makers

To get started, upload the **Final Project.zip** file (which is available in the Final Project Folder in Blackboard) to your RStudio cloud. The .zip file includes the dataset (saved as an RDS data frame file) and the R notebook template which sets everything up for you.

loan_data (R data frame)

The *loan_data* data frame contains information on 3-year loans that were originated in 2013 by a local bank for customers residing in the United States. The company is looking to see if it can determine the factors that lead to loan default and whether it can predict if a customer will eventually default on their loan at time of loan origination. The goal is to become better at identifying customers at risk of defaulting on their loans to minimize the bank's financial losses.

The dataset contains a mixture of applicant demographics (gender, age, residence, etc..), financial information (income, debt ratios, FICO scores, etc..), and applicant behavior (number of open accounts, historical engagement with the bank's products, number of missed payments, etc...)

Variable	Description
loan_default	Did the loan result in default (Yes/No)
residence_property	Applicant residence status - Rent vs Ownership
gender	Applicant gender (Female/Male)
age_category	Applicant Age
highest_ed_level	Applicant highest education level – Less than High School through PhD/Doctorate degree
us_region_residence	Applicant US region of residence
loan_amnt	The loan amount in USD
adjusted_annual_inc	Applicant adjusted annual salary (for local cost of living) in USD
pct_loan_income	The loan amount as a proportion of adjusted income
fico_score	Applicant FICO score (ranges between 300 and 850)
dti	Applicant debt to income ratio
inq_last_6mths	Number of credit inquires in the last 6 months
open_acc	Applicant total open accounts at bank
bc_util	Applicant bank utility score – a measure of applicant engagement with the bank's products
num_accts_ever_120_pd	Number of loan accounts that were ever 120 days past due for this applicant
pub_rec_bankruptcies	Applicant public bankruptcies

Executives at this bank have hired you as a data science consultant to identify the factors that lead to loan default and to recommend data mining algorithms that can predict whether a customer will default on their loan.

Specifically, the broad questions that the bank is trying to answer include:

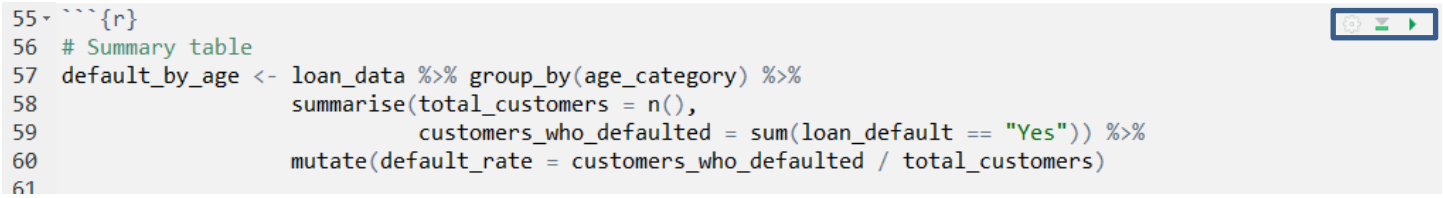
1. What are the factors that contribute to customers defaulting on their loans?
2. Is it possible to predict whether a customer will default on their loan? If so, how accurate are the predictions?
3. How many costly errors does the predictive model produce (customers classified as not defaulting, but eventually do)?

R Notebook File (10 Points)

The **Final Project.zip** file contains the **R Notebook Template for Final Project.rmd** file for this project. Upload the Final Project.zip file to your RStudio Cloud to access the file and the project data set. You will be writing your **R** code and analysis in this file. **R** notebooks are an easy way to communicate results from data mining projects because you can write **R** code in special sections called chunks, followed by plain text outside of chunks for comments or analysis.

Any analysis that you need to write can be typed in this file ***outside*** of the **R** code chunks. Your **R** code needs to be placed within code chunks that you can create by typing "Ctrl" + "Alt" + "I" or by selecting "**R**" under the Insert button on the top right of the file in RStudio.

As you build your final project code and analysis, you can check if your **R** code is working by clicking on the "play" button within each chunk



```
55 ```{r}
56 # Summary table
57 default_by_age <- loan_data %>% group_by(age_category) %>%
58                   summarise(total_customers = n(),
59                             customers_who_defaulted = sum(loan_default == "Yes")) %>%
60                   mutate(default_rate = customers_who_defaulted / total_customers)
61
```

To get full credit for this section, you must:

- submit your final **R** notebook file with your **R** code and analysis titled as **MIS 431 Final Project YOUR NAME(S)**
- Have comments within your R code chucks explaining what the code is doing at a high level
- Include the name and G Number of ***all*** students working on the project (up to 2 students per group)

Exploratory Data Analysis and Variable Importance Analysis (40 Points)

This portion will represent the first two sections of your **R** notebook file.

1. Exploratory Data Analysis (EDA) 30 Points
    a. You must think of at least **8 relevant questions** that explore the relationship between ***loan_default*** and the other variables in the data set
        i. An example would be "Do loan default rates differ by customer age?"
    b. ***You must answer each question*** and provide supporting data summaries with either a summary data frame (using ***dplyr***) or a plot (using ***ggplot2***) or both
    c. In total, you must have **at least 5 plots and 5 summary data frames** for your EDA section
    d. You must use **at least 3 different plot types** (ex. bar chart, histogram, heat map)

2. Variable Importance and Selection (On the Training Dataset Only!) 10 Points
    a. In this section you must explore which variables are important for predicting ***loan_default*** (the response variable for classification)
    b. You must perform ***variable selection*** on the training data using the following method:
        - Variable Importance Analysis using random forests
    c. Include a ***short analysis of your results and justification for the final variables you have chosen to keep in your final predictive model***.

Predictive Modeling (40 Points)

This portion will represent the third, fourth, and fifth sections of your R notebook file. For **all** the predictive models that you try, you must use ***cross validation*** to study prediction accuracy. This means you must split your original data into a training and test set (this has been done for you in the Notebook template file), and perform the necessary model building steps as outlined in the course lectures and tutorials.

1. Classification – Predicting loan_default
    a. Try **three** classification algorithms to predict the response variable, ***loan_default***, on the training data
        i. Algorithms you can try include:
            1. Logistic Regression
            2. LDA
            3. QDA
            4. Naïve Bayes
            5. Decision Trees
            6. Random Forests
            7. KNN
    b. Analyze the results from each model
        i. Use the **cf_matrix** function to analyze the various error rates on the training data and choose an optimal probability cut-off based on this analysis. Write a short analysis of your ***false positive rates***, ***false negative rates,*** and F_1 ***scores*** for both models to justify your choice.
        ii. Using the optimal probability cut-off values from part (i), make predictions on the test data and provide a summary of the confusion matrix and discuss model accuracy in terms of the various error rates on the test dataset. Remember that when speaking about model accuracy, you must always apply your trained model to a test dataset which was not used in ***any*** of the model building steps.

Summary of Results and Recommendations to Decision Makers (10 Points)

Write a summary of your overall findings and recommendations to the executives at the bank. Think of this section as your closing remarks of a presentation, where you summarize your key findings, model performance, and make recommendations to improve loan processes at the bank. Your summary should include:

1. Key findings from your EDA and Variable Importance analysis. What were the things that stuck out for you in this section and why are they important?
2. Your "best" classification model and a confusion matrix analysis for this model, including a discussion of either the false negative rate or false positive rate (which ever you think is more important to guard against)
3. Your recommendations to the bank – how could loan default rates be improved?