Final Project

Objective

The goal of the final project is for you to apply what you have been learning in this course to address a real-world business data mining problem. Four business cases have been provided to you, which include business problem background, data, and variable descriptions. In your groups, you will choose a business case and perform **at least** 2 types of analysis. **At least one of your analyses must be a classification method.** All analysis should be completed using the same business case and dataset. You will produce a written report based on your analysis, which should describe your analysis and key findings in a clear, concise manner to managers of the business described in the business case. During the Week 10 class session you will present your findings to the class.

Analysis Methods

Unsupervised Methods	**Supervised Methods (Classification)**
Cluster Analysis (kMeans, Hierarchical)	Logistic Regression
Association Analysis	Naïve Bayes
Principal Components Analysis (PCA)	k-Nearest Neighbors
	Support Vector Machines
	Decision Trees
	Ensemble Methods
	Artificial Neural Networks

Deliverables

1. Project Proposal

-**Due 8/1 at 11:59 PM**

-Project Proposal should be submitted as a PDF (approximately 1 page)

-Project proposal should include:

-Group #, Name(s)

-Chosen case study or alternate data and business problem chosen (and data link, variable descriptions)

-Motivation for choosing the business case/data

-Initial data overview (quality, dimensionality, identification of target variable for classification, identification of variable types (numerical, categorical (ordinal/nominal))

-Initial exploratory data analysis results

2. Presentation

-In class on 8/23

-PPT slides in PDF format due on 8/23 at 11:00 AM)

-Your presentation should be 5 minutes

-Powerpoint slides should be used.

-All group members must present. To receive credit for the presentation, each member must participate in the presentation. If you are absent, you will receive a 0 (not the group's presentation grade).
-The written report should guide the content of your presentation
-The intended audience is managerial. Focus on high-level information and insights, including:
-A description of your data, business problem/objectives and the relevance/importance of your project
-What are you trying to predict or understand using the data? What kind of analyses are you doing?
-Present your findings, including performance, goodness of fit and validation information.

3. Report

-**Due 8/29 at 11:59 PM**
-Report should be submitted as a PDF File
-File naming should be: Final_Group#.PDF
-The written report should describe your analysis and key findings in a clear, concise manner to a business audience.
-Note: The written report should **not** contain any R code.

The report should include:

A. Title, Group Number, Name(s)
B. Introduction
    a. Introduce your data, motivation and objectives.
C. Data
    a. Describe your data sample (dimensionality, quality, etc.).
    b. Describe your variables (numerical, nominal, ordinal).
    c. Include descriptive statistics and exploratory data analysis and visualizations.
    d. Overview of data pre-processing, cleansing and transformation for analysis.
D. Analysis Results
    a. Present your analysis findings in-depth.
    b. Summarize why you chose your analysis method, the analysis itself (including any modeling decisions made) and the results.
    c. Present internal or external validation measures and accompanying plots, as necessary.
E. Discussion & Conclusion
    a. Discuss the high-level findings of your analysis in words.
    b. Discuss the implications of your findings for the business in the case study.
    c. Discuss how your analysis addresses or solves the business problem.
    d. Conclude by describing next steps that the business can take based on your findings.
F. References
    a. You must include a link to your data in the references of your report.
    b. Any resources/sources used should be appropriately cited in-text and/or in a references section. Citations should be in MLA or APA format.

Note: You can structure the report differently, but these major elements and the listed required content should be incorporated into your report. Report should be written in paragraph format and should not include bullets.

4. .R Script File

-**Due 8/29 at 11:59 PM**

-File naming should be: Final_Group#.R

-For any random seed initialization, you should use your birthday seed. If you are working individually, it will be MMDD. For example, my birthday seed is 831, since my birthday is August 31st. If you are working in a group of 2, your seed should be MMDDMMDD. For example, if my group member's birthday is March 6, my random seed would be 83136 for all procedures requiring random seed initialization. Note: the maximum number of digits is 9. If your group's seed exceed 9 digits, truncate the seed number.

-All code should be original, based on the class materials and not copied from any external source.

-Code should be appropriately commented/annotated

Note: The purpose of the project is to demonstrate an understanding of the materials covered in the course. As such, code copied from the internet and submitted as your own, original work will receive a Final Project grade of 0. Short snippets of code, however, can be used with proper citation. Class code does not require citation.

5. .RData File

-**Due 8/29 at 11:59 PM**

-File naming should be: Final_Group#.RData

-File must contain all objects, values and data used to complete your project

-Points will be deducted if I cannot sequentially replicate your analysis using the .R and .RData files

Business Cases

Case #1: Predicting Productive Audits

Background: An accounting firm would like to use data about their clients to help them make auditing decisions. Their dataset represents their clients demographic and tax return information. It is costly to audit a client when an audit is not necessary, so they need to strategically predict which clients' audits will be productive.

Data File: CompanyAudit.csv

Variable Descriptions:

ID: Unique identifier for each person.
Age: Age of person.
Employment: Type of employment.
Education: Highest level of education.
Marital: Current marital status.
Occupation: Type of occupation.
Income: Amount of income declared.
Gender: Gender of person.
Deductions: Total amount of expenses that a person claims in their financial statement.
Hours: Average hours worked on a weekly basis.
Audit_Type: indicates nonproductive and productive audits, respectively. Productive audits (1) are those that result in an adjustment being made to a client's financial statement and nonproductive audits (0) are those that do not.

Case #2: Predicting Credit Card Defaults

Background: A credit card company wants to use their data on existing customers to help them make informed decisions about which future customers to offer credit to. The credit card company has data including customer demographics, credit, and payment information. They want to be able to predict if a customer will default. It is very costly to the company to lose business because they believe a good customer will default and to give a customer credit who ends up defaulting.

Data File: CCDefault.csv

Variable Descriptions:

ID: Unique identifier of customer
LIMIT_BALANCE: Amount of the given credit, including both the individual consumer credit and family (supplementary) credit
GENDER: 1 = male; 2 = female
EDUCATION: 1 = graduate school; 2 = university; 3 = high school; 4 = others
MARITAL_STATUS: 1 = married; 2 = single; 3 = others
AGE: Age in years.
PAY_1 – PAY_6: History of past payment (PAY_1 = September, PAY_6 = April) . Represents the payment records of the customer (from April to September, 2005) -1 = on-time payment; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above. – 2 indicates that the customer did not have any consumption and 0 indicates the use of revolving credit.
BILL_AMT1 - BILL_AMT6: Amount of monthly bill (BILL_AMT1 = September, BILL_AMT6 = April)
PAY_AMT1 - PAY_AMT6: Amount paid of monthly bill (PAY_AMT1 = September, PAY_AMT6 = April)
DEFAULT: Indicates if the customer's credit card is in default (1) or not (0)

Case #3: Predicting Customer Churn

Background: Customers leave their telecommunications company frequently because they are dissatisfied. Even with low percentages of customer churn, companies can lose millions of dollars a month due to customers leaving. For this reason, a telecommunications company would like to use their data on past and present customers to help them to better understand their customers and predict if a customer will leave the company. It is costly to the company to lose a good customer, so they want to minimize cases where they predict a customer will stay and they actually leave.

Data File: CustomerChurn.csv

Variable Descriptions:

customerID: unique customer identifier
gender: gender of customer,
SeniorCitizen: indicates if a customer is a senior citizen (1) or not (0)
Partner: Indicates if the customer has a partner (Yes) or not (No)
Dependents: Indicates if the customer has dependents (Yes) or not (No)
tenure: the length of time that the customer has been a customer
PhoneService: Indicates if the customer has phone service with the company (Yes) or not (No)
InternetService: Indicates if the customer has fiber optic, DSL or no internet service with the company
Contract: The type of contract that the customer has with the company (Month-to-month, One year, Two year)
PaperlessBilling: If the customer is enrolled in paperless billing (Yes) or not (No)
PaymentMethod: The most recent payment method used by the customer to pay the company (Electronic check, Mailed check, Bank transfer (automatic), or Credit card (automatic)
MonthlyCharges: The most recent amount that the customer is charged per month
TotalCharges: The total amount that the customer has been charged
Churn: Whether the customer has left the company (Yes) or not (No)

Case #4: Predicting Wine Quality

Background: A high-end beer and wine distributor wants to better understand their product offerings and automate the process of choosing high quality wines to distribute. They have data about their existing products regarding the chemical makeup and color and each wine is identified based on its quality. The distributor wants to be able to correctly predict high quality wines that they would like to sell. It is costly to the company to predict that a wine is high quality when it is actually low quality.

Data File: WQMarketing.csv

Variable Descriptions:

fixed_acidity
volatile_acidity
citric_acid
residual_sugar
chlorides
free_sulfur_dioxide
total_sulfur_dioxide
density
pH
sulphates
alcohol
color: wine color, either red or white
quality: quality level based on sensory data, either High or Low