CIND 119 Class Project

by Syed Shariyar Murtaza, Ph.D.

Project Description

Your task in this project is to work with a team of data scientists and solve a problem based on a given dataset for an organization. There are three datasets available for this project. These datasets are described at the end. You will have to choose **only one** of the dataset for analysis for your project. Your team will consist of up to a maximum of 2 students and each team will submit one project report. In your project, you will perform the following steps on the selected dataset:

1. Data Preparation.
2. Predictive Modeling/Classification:
   a. Classification using Decision Tree.
   b. Classification using Naive Bayes.
   c. Compare the results of the 2 techniques on original and filtered data.
3. Conclusions and Recommendations

Describe your results in your project report for each of the above steps. These steps are further explained in the sections below. You can use any tool of your choice for this project, such as SAS, Python, R, Weka, etc. **Note that the final project report to be submitted will not include any information about code or the tools (specially no screenshots of tools) but the actual results and description of the analysis in the form of text, tables and charts.**

Data Preparation

The first and foremost step of data mining process is to understand the data and identify the research question(s). Below are some suggestions to explore and understand datasets:

- Look at the attribute type; e.g., nominal, ordinal or quantitative.
- Find any missing values.
- Find max, min, mean and standard deviation of attributes.
- Determine any outlier values (records) for each of the attributes or attributes under consideration (min, max, std. dev, scatter plots, box plots or others can be used).
- Analyze the distribution of numeric attributes (normal or other). Plot histograms for attributes of concern and analyze whether they have any influence on the class attribute.
- Which attributes seem to be correlated? Which attributes seem to be most linked to the class attribute?.
- Which attributes do you think can be eliminated or included in the analysis? This can be a subjective decision or an objective decision based on a statistical method.

- Determine whether the dataset has an imbalanced class distribution (same proportion of records of different types or not) and do you need to balance the dataset.
- Determine whether you need to handle missing values or transform any attributes (e.g., by normalizing the attributes, discretizing numeric attributes to categorical attributes, etc.).

Note that it is not necessary to perform all of the above steps, you can decide to perform one or more steps as you deem appropriate. For example, if you judge that there are no outliers in a dataset then you do not need to apply any method to remove outliers. Similarly, if there are no missing values then there is no technique needed to fix them, or avoid any step that is too difficult to implement for you. Finally, describe your findings for data preparation in your report.

Predictive Modeling (Classification)

Apply the classification algorithms, Decision Tree and Naïve Bayes, which you studied in the course on your dataset after data preparation phase.

- You will predict the class attribute by using each classification algorithm.
- Determine the right strategy for dataset split: simple train-test set split, or 10-fold cross validation. *Choose only one strategy; e.g., train-test set split.*
- Understand the output of your algorithms as much as possible; e.g., which attributes are used by decision tree to make the decision.
- Determine your performance measures (accuracy, true positive, false positive, etc.).
- Identify which algorithm performs well out of the two by using true positive, false positive and accuracy measure.

Describe results of predictive modeling in your report. Explain your interpretation of output of each of the algorithm in your report (e.g., explain decision tree). Explain how you determined the best performing algorithm.

How to Compare Your Classification Models

In order to evaluate and compare machine learning models with different features, a known approach is to create a baseline model first. This can be done by training a classification algorithm (e.g., decision tree) on the training set using the entire feature set (all attributes) and evaluating its performance using the selected metric (such as accuracy, true positive rate, and false positive rate) on the test set. Second, you need to train the same classification algorithm on your selected features and evaluate its performance on the validation set (or test set) using the same selected metrics as before. Similarly, you will repeat these two steps for the second classification algorithm—Naïve Bayes. **Finally, you would compare the performance of different trained models by using the selected metrics.** This would give an indication whether your selected features (or parameters) have increased or decreased the performance of trained models.

Conclusion and Recommendations

State your major findings from different sections. State your recommendation to the company that they can put into place to solve their problem.

Tools

Two Python notebooks are attached with this project description. They will serve as a guide for you on how to analyze these datasets using Python. Two tutorials use two different types of Python packages to analyze data. Both tutorials are only tested using Google's Colab, it is recommended to run them on Colab only. Upload the given notebook files (.ipynb) on Colab and start experimenting.
You can also use SAS of this data analysis using some of the labs we have provided you earlier in the course.
If you are taking other courses in this program, then R can be use as well.
You can also use GUI tools like Weka. Tutorial 2 uses its Python package. You can also use it by installing as a software application (https://www.youtube.com/watch?v=TF1yh5PKaqI)
In short, choice of tools is yours.

Datasets

Select only one of the following datasets for your group and solve the problem for this company described for a particular dataset below.

Dataset for Churn

Customer churn or customer attrition means the loss of customers for a company. The problem for customer churn is that the company would like to know in advance which customers would churn in near future. You are a member of a team of data scientists and the task of your team is to help this company in characterizing customer churn through data analytics methods. This dataset has 21 attributes including a binary class attribute about churn. The descriptions of the attributes are given below:

1. State: Customer's state.
2. Account Length: Integer number showing the duration of activity for customer account.
3. Area Code: Area code of customer.
4. Phone Number: Phone number of customer.
5. Inter Plan: Binary indicator showing whether the customer has international calling plan.
6. VoiceMail Plan: Indicator of voice mail plan.
7. No of Vmail Mesgs: The number of voicemail messages.
8. Total Day Min: The number of minutes the customer used the service during day time (continuous quantitative data type).
9. Total Day Calls: Discrete attribute indicating the total number of calls during day time.
10. Total Day Charge: Charges for using the service during day time (continuous data type).

11. Total Evening Min: The number of minutes the customer used the service during evening time.
12. Total Evening Calls: The number of calls during evening time.
13. Total Evening Charge: Charges for using the service during evening time.
14. Total Night Min: Number of minutes the customer used the service during night time.
15. Total Night Calls: The number of calls during night time.
16. Total Night Charge: Charges for using the service during night time.
17. Total Int Min: Number of minutes the customer used the service to make international calls.
18. Total Int Calls: The number of international calls.
19. Total Int Charge: Charges for international calls.
20. No of Calls Customer Service: The number of calls to customer support service.
21. Churn: Class attribute with binary values (True for churn and False for not churn).

Bank Marketing Dataset

This dataset refers to the problem of telemarketing for a bank. The dataset is collected from a Portuguese bank and the bank wants to have an effective telemarketing strategy to sell long-term deposit accounts (e.g., bonds, saving accounts, etc.). These marketing campaigns were based on phone calls and multiple contacts were often needed to determine whether a customer would subscribe to a long-term deposit account. Your team of data scientists will help this bank in determining such customers and devising an effective telemarketing strategy by applying data analytics method on the given dataset.

1 – Age: Age of the customer (numeric).
2 - Job: Type of job (qualitative).
3 - Marital: Marital status (qualitative).
4 – Education: Education of the customer (qualitative).
5 - Default: Shows whether the customer has credit in default or not (qualitative).
6 - Balance: Average yearly balance in Euros (numeric).
7 - Housing: Shows whether the customer has housing loan or not (qualitative).
8 - Loan: Shows whether the customer has personal loan or not (qualitative/categorical).
9 - Contact: Shows how the last contact for marketing campaign has been made (qualitative)
10 - Day: Shows on which day of the month last time customer was contacted (numeric).
11 - Month: Shows on which month of the year last time customer was contacted (qualitative).
12 - Duration: Shows the last contact duration in seconds (numeric).
13 - Campaign: Number of contacts performed during the marketing campaign and for this customer (numeric).
14 - Pdays: Number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted).
15 - Previous: Number of contacts performed before this campaign and for this client (numeric).
16 – Poutcome: Outcome of the previous marketing campaign (qualitative).

17 - Y – Class attribute showing whether the client has subscribed a term deposit or not (binary: "yes","no")

Credit Card Dataset

In order to provide loans to customers, a bank needs to make right decision in determining who should get the approval and who should not. This dataset is the German Credit Data that contains 20 attributes and the class attribute showing a good or a bad credit risk. Your team of data scientists will need to develop a data analytics based strategy for the bank managers that can help them in making a decision about loan approval for the prospective applicants/customers.

1. Creditability: The class attribute (qualitative)showing whether the credit rating is good or bad.
2. Account Balance: Checking account status (1: < 0 DM, 2: 0<=...<200 DM, 2 > 200 DM, 4: No checking account), where DM= Deutsche Mark (qualitative attribute).
3. Duration of Credit (month): Duration of credit in months (numerical)
4. Payment Status of Previous Credit: Credit history (qualitative) 0: no credits taken, 1: all credits at this bank paid back duly, 2: existing credits paid back duly till now, 3: delay in paying off in the past, 4: critical account.
5. Purpose: Qualitative attribute showing the purpose of the loan (0: New car, 1: Used car , 2: Furniture/Equipment, 3: Radio/Television, 4: Domestic Appliances , 5: Repairs ,6: Education ,7: Vacation, 8: Retraining ,9: Business, 10: Others)
6. Credit Amount: Numerical value showing the credit amount
7. Value Savings/Stocks: Qualitative attribute showing average balance in savings and stocks (1 : < 100 DM, 2: 100<= ... < 500 DM, 3 : 500<= ... < 1000 DM, 4 : =>1000 DM, 5: unknown/ no savings account)
8. Length of current employment: Qualitative attribute showing length of employment (1 : unemployed, 2: < 1 year, 3: 1<=...<4 years, 4: 4<=...<7 years, 5:>=7years).
9. Instalment percent: Installment rate in percentage of disposable income (numerical)
10. Sex & Marital Status: Qualitative attribute showing gender and marital status (1: male : divorced/separated, 2: female : divorced/separated/married, 3 : male: single, 4: male : married/widowed, 5 : female : single)

11. Guarantors: (Qualitative) Guarantors and co-applicants: (1 : none, 2 : co-applicant, 3 : guarantor)
12. Duration in Current address: Qualitative value showing the duration in current address (1: <= 1 year, 1<...<=2 years, 2<...<=3 years, 3:>4years)
13. Most valuable available asset: Qualitative attribute showing valuable assets (1 : real estate 2 : savings agreement/ life insurance, 3 : car or other, 4 : unknown / no property)
14. Age (years): Numerical value showing age in years.
15. Concurrent Credits: Installment plans (1 : bank, 2 : stores, 3 : none)
16. Type of apartment: Type of housing (1 : rent, 2 : own, 3 : for free)
17. No of Credits at this Bank: Numerical value showing number of existing credits at the bank
18. Occupation: Job (Qualitative) (1 : unemployed/ unskilled - non-resident, 2 : unskilled - resident, 3 : skilled employee / official, 4 : management/ self-employed/highly qualified employee/ officer)
19. No of dependents: Numerical value showing number of dependents
20. Telephone: Qualitative attribute for telephone number (1: yes, 2: No)

21. Foreign Worker: Qualitative attribute showing whether the person is the foreign worker or not (1: yes , 2: no)

Project's Evaluation Criteria

Each group will be evaluated based on its analysis of their project. All the individuals will be evaluated based on workload distributions among the groups. Please use the project template to submit one group report including workload distribution. Five percent bonus marks if you can think of creative ways of using SQL/NoSQL/Tableau with data of this project.

Summary/abstract	20% marks
Data preparation	35% marks
Predictive Modeling	35%marks
Conclusions and Recommendation	10% marks
Bonus marks for creativity	5% marks