INFO 659 Data Mining Application

**Assignment 3** (10 points) Data Mining Report on

Data Transformation, Models, and Evaluation

# Objective and Setup

This assignment is to follow the preliminary work of Assignment 2 on predicting credit card default and conduct a more formal analysis with data transformation, modeling, and evaluation.

You can **reuse code from Assignment 2**. In addition, you should **review your code for week 6 and week 7** **exercises** on transformation, modeling, and evaluation.

Again, the data is given in the CSV format, available at (same data as in Assignment 2):

<https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset>

The major task of this assignment is to build models to predict credit card defaults based on **identified relations of variables** in the data, and conduct evaluation.

# Tasks

## **Understanding variables and relations in data (2 points)**

**A.1.** Discuss how **credit and payment history data** such as PAY\_AMT1 have an impact on payment default.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Data Type** | **Potential impact on “Default” and reason** |
| Limit\_Bal |  |  |
| Pay\_0, 2, 3, 4, 5, 6 |  |  |
| Bill\_Amt1, 2, 3, 4, 5, 6 |  |  |
| Pay\_Amt1, 2, 3, 4, 5, 6 |  |  |

**A.2.** Discuss in what ways some of the **above attributes** contribute to default.payment.next.month **together**. **Please identify at least two pairs** of attributes that can be treated together and how.

|  |  |  |
| --- | --- | --- |
| **Variable 1** | **Variable 2** | **Discuss their relation, how to combine them (ratio, difference, or others) and your reason/theory** |
|  |  |  |
|  |  |  |
|  |  |  |

Note it is possible that *more than two variables* can be taken together. If that is the case, extend the above table to accommodate more than two variables. You may review **week 7 tutorials on transforming web filter** data for related ideas.

## **Data preparation and cleansing (1 points)**

B.1.Load data and initial data conversion/transformation:

1. Load“UCI\_Credit\_Card.csv” into **data frame** variable in R using read.csv().
2. Convert the following variables into as nominal (categorical, factor) variables: **Sex, Education, Marriage, Pay\_ ?**and **default.payment.next.month**.

3) Use class() function check on Sex, Education, Marriage, Pay\_? and default.payment.next.month, they should ALL be **“factor”** variables.

B.2. Create a filtered dataset with only **non-negative amounts**.

1. Use the subset() function to select only positive values on the **6 BILL\_AMT variables** and **6 PAY\_AMT variables.** Like (fill the … with actual:

ccnn <- **subset**(cc, BILL\_AMT1>=0 & BILL\_AMT2>=0 & PAY\_AMT1>=0 & PAY\_AMT2>=0 & …)

nrow(ccnn)

1. Check the number of rows in the filtered subset and you can use View(ccpo) to double check on the data.

## **Data Transformation and Classification/Modeling (4 points)**

**C.1.** Pick **one classification metho**d, model with **default.payment.next.month ~** **variables in A.1.**, and evaluate:

* You can pick one of these methods: *Naïve Bayes*, *Decision Tree*, *SVM*, and *Neural Networks*, *Logistic Regression*, among others we have discussed.
* Select 90% of data for training and 10% for testing;
* Build a model with training data (90% data) to predict *default.payment.next.month*, using **at least three variables from A.1.**
* Run prediction with the model on test data (10% data) and **record the following scores**:
  + Present the confusion table with **TP, TN, FP,** and **FN**
  + Report **Accuracy, Precision, Recall, F,** and **Kappa** in **Table D.**

**C.2.** Perform **data transformation (with new relational attributes)** and redo classification:

1. Follow the **treatments (at least two relations)** of the variable pairs you have identified in **A.2**;
2. Create **new variables** that compute the relations you have identified in **A.2**;
3. Build a model with training data (90% data) to predict *default.payment.next.month*, using the **new relational attributes** (**plus any other variables** you would like to include) here.
4. Run prediction with the model on test data (10% data) and **record the following scores**:
   1. Present the confusion table with **TP, TN, FP,** and **FN**
   2. Report **Accuracy, Precision, Recall, F,** and **Kappa** in **Table D.**

**C.3.** Examine attribute **value distribution** (histogram),and **perform log transformation** on attributes you see fit:

1. Create a **new attribute** that is the logarithm **of each attribute** with an extremely wide, **“skew” distribution**.
2. Remove attributes that are no longer needed in your analysis.
3. Hopefully data distributions now look **“normal”**.
4. Build a model with training data (90% data) to predict *default.payment.next.month*, using at the **new relational** (and log-transformed) **attributes** **plus any other variables** you would like to include here.
5. Run prediction with the model on test data (10% data) and **record the following scores**:
   1. Present the confusion table with **TP, TN, FP,** and **FN**
   2. Report **Accuracy, Precision, Recall, F,** and **Kappa** in **Table D.**

Here is an example of transforming a “skew” distribution to a “normal” distribution based on logarithm:

|  |  |
| --- | --- |
| Macintosh HD:Users:weimao:Desktop:Screen Shot 2017-11-14 at 4.42.24 PM.png | Macintosh HD:Users:weimao:Desktop:Screen Shot 2017-11-14 at 4.43.41 PM.png |
| **Before** log transformation | **After** log-transformation |

**C.4.** Pick **another classification model** or the same model with **different parameter values**, and repeat the modeling and evaluation as in **C.3**. Report the confusion table and results to **Table D.**

## Evaluation and Results (2 points)

**D.** Evaluation Results of Different Models:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Method | **C.1.** Classification **without** Transformation | | | | |
| Correct % | Precision | Recall | F | Kappa |
| C.1 | Model 1 Name, Variables, Parameters, etc. |  |  |  |  |  |
| C.2 | Model 1 Name, Variables, Parameters, etc. |  |  |  |  |  |
| C.3 | Model 1 Name, Variables, Parameters, etc. |  |  |  |  |  |
| C.4 | Model 2 Name, Variables, Parameters, etc. |  |  |  |  |  |

## Report with Interpretation and Conclusion (3 points)

Discuss the results in Task D and answer the following questions:

**E.1.** In terms of the **reasons and theories** presented in tasks A1 through A2, which ones have been confirmed by your analysis? Please discuss even if there is no obvious answer.

**E.2.** Does **data transformation (with new relational variables in C.2)** help? Which one helps most and why? Or which does not?

**E.3.** Which classification **method**(s) and/or **parameters** appear to perform well? Which ones do not?

**E.4.** Reviewing results in Task D, which **evaluation metrics** (of Correct%, Kappa, F, Precision, and Recall) best capture how good/poor the result is? Which metric is not as helpful?

**E.5.** Pick the most helpful evaluation metric, which method (with what data transformation if applicable) is the **overall winner** of the results? Reason about why the method performs well.

**Submit the report (data mining notebook) to Blackboard.**