CENTER FOR INFORMATION SYSTEMS AND TECHNOLOGY

**IST 340** – **Knowledge Discovery and Data Mining**

# Description of Final Project

**1. INTRODUCTION (Classification Problem)**

Your Data Mining Project Team is requested to develop a predictive model to determine characteristics of students who were good loan risks and those that were poor loan risks. The target variable (i.e. *NoPaymentDue*) records whether the given student was behind on his/her payments one (1) calendar year after he/she was due to start making payments on the loan. For a given student, the input variables represent the given student’s situation and history at the time that he/she applied for the loan. Table 1 below displays the Input and Target variables for the cases that should be used to build the predictive model. The data for each case is obtained from multiple data files (see Table 2). The relationships between fields in Table 1 and 2 are described in Table 3.

Consider you and your team members to be the Data Mining Analytical Experts on this project team. For each Task that is described in the next section you will have to do the following:

1. Develop and describe your plan for addressing the task that includes a detailed description of each step in the appropriate order. The description of your plan should not include material that you obtained from the execution of your plan! You must develop & document your plan before you execute it.
2. Execute your plan. Provide evidence of your work in a manner similar to that required for CE1.
3. Create a Score dataset of 35 rows that includes all variables except the target variable. For each task, apply the ‘best’ model to score this dataset.

**Table 1: Variable Description**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **PossibleModel Role** | **Measurement** |
| StudentID | ID | Interval |
| NoPaymentDue | Target | Binary |
| Enlisted | Input | Binary |
| Service | Input | Nominal |
| Disabled | Input | Binary |
| Unemployed | Input | Binary |
| Enrolled | Input | Binary |
| School | Input | Nominal |
| Region | Input | Nominal |
| Country | Input | Nominal |
| Longest\_Absence\_From\_School | Input | Interval |
| Gender | Input | Binary |
| Marital\_Status | Input | Binary |
| Filed\_for\_Bankrupcy | Input | Binary |
| LongTermPlanningScore | Input | Interval |
| DecisionStyle | Input | Nominal |
| PowerOrientedScore | Input | Interval |
| CommunityOrientedScore | Input | Interval |
| Family\_Income | Input | Interval |
| Parents’ Education Level | Input | Ordinal |
| Credit Score\_Raw | Input | Interval |
| Credit Score\_AgeAdj | Input | Interval |
| Loan Amount | Input | Interval |
| HS\_Math | Input | Ordinal |
| HS\_Science | Input | Ordinal |
| HS\_English | Input | Ordinal |

**Table 2: Data Sources**

|  |  |
| --- | --- |
| **FileName** | **Variable(s) Name** |
| No\_payment\_due | StudentId, NoPaymentDue (i.e. pos, neg) |
| Male | StudentId |
| Region, Marital Status xls (Spreadsheet) | StudentId, Country, Region, Marital\_Status |
| Longest\_Absense\_From\_School | StudentId, Number\_of\_Months |
| Family\_Income xls | StudentId, Family\_Income |
| Enrolled | StudentId, School, Units |
| Enlist | StudentId, Service |
| Unemployed | StudentId |
| Filed\_for\_Bankrupcy | StudentId |
| Disabled | StudentId |
| Parents\_Education.xls | StudentId, Parents’ Education Level |
| LoanAmount.xlsx | StudentId, Loan |
| CreditRating Data.xlsx | StudentId, Credit\_Score\_Raw, Credit\_Score\_AgeAdj |
| HS\_Academics | StudentID, HS\_Math, HS\_Science, HS\_English |
| PersonalityCharacteristics.xlsx | StudentID, LongTermPlannerScore, DecisionStyle, PowerOrientedScore, CommunityOrientedScore |

**Table 3: Relationships between Variables & Data Sources**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Source** | Value |
| StudentId | No\_payment\_due: StudentId |  |
| NoPaymentDue | No\_payment\_due: NoPaymentDue |  |
| Enlisted | Enlist | **1** if the corresponding StudentId is in the **Enlist** file; **0** otherwise |
| Service | Enlist: Organization |  |
| Enrolled |  | **1** if the corresponding StudentId is in the **Enrolled** file; **0** otherwise |
| School | Enrolled: School |  |
| Units | Enrolled: Units |  |
| Disabled | Disabled: | **1** if the corresponding StudentId is in the **Disabled** file; **0** otherwise |
| Unemployed |  | **1** if the corresponding StudentId is in the **Unemployed** file; **0** otherwise |
| Filed\_for-Bankrupcy |  |  |
| LongestAbsence | Longest\_absense\_from\_school: Number\_of\_Months |  |
| Male |  | **1** if the corresponding StudentId is in the **Male** file; **0** otherwise |
| Marital Status | Region, Marital Status xls | **1:** Married  **0:** Single |
| Region | Region, Marital Status xls | **1, 2, 3, 4, 5** |
| Parents’ Education Level | Parents\_Education.xls | ‘VERY LOW’, ‘LOW’, ‘MEDIUM’, ‘HIGH’, ‘VERY HIGH’ |

**2. DESCRIPTION OF TASKS**

**Task A1:**

Initially, the end-users simply expressed that a predictive model that was very accurate would have a significant impact on their operations. The project leader, Ms. Imani Kuhn, who had done some graduate work in Management Science, also told the project team that before the model was released for use by the end users that they had to persuade her that the recommended model was indeed the ‘best’ model. You should consider the use of derived variables. The target event is *NoPaymentDue*  = ‘pos’.

**Task A2:**

Some end-user members of the project team stated that while they were pleased with the accuracy of the model of task A1, they thought that it would be better to have an explanatory model. It was also felt that if the explanatory predictive model is a DT then ideally it should have between 4 to 6 rules; and that DTs with more than 9 rules and less than 3 rules would be of no value. However, other types of explanatory models should also be considered. Given these concerns, you were then requested to generate the most appropriate predictive model. You should consider the use of derived variables. The target event is *NoPaymentDue* = ‘pos’.

**Task A3:**

After reviewing the outputs of the previous tasks, Ms. Imani Kuhn suggested that it might be useful to also look at the most appropriate predictive model that was focused on obtaining the highest lift for the top 30% of the cases. You should consider the use of derived variables. The target event is *NoPaymentDue* = ‘pos’.

**Task A4:**

After reviewing the outputs of the previous tasks, some of the end-user experts remarked that it might also be useful to incorporate the fact that it was more costly to misclassify a person from whom a payment was really due than to misclassify a person from whom no payment was due. Ms. Imani Kuhn suggested that they should research the relevant costs. The result of their research was that there was an average profit of $817.00 for correctly classifying a person for whom a payment was really due while there was an average profit of $245.00 for correctly classifying a person for whom no payment was really due. On the other hand, there would an average loss of $671.00 for misclassifying a person for whom a payment was really due while there was an average loss of $465.00 for misclassifying a person for whom no payment was really due. The target event is *NoPaymentDue* = ‘pos’.

**Task A5:**

Mr. David Braud, an end-user, has recently been exposed to the concept of clustering. He suggested to Ms. Kuhn that it might be useful to see if there were three (3) natural groups of students based on the interval and ordinal input variables. Ms. Kuhn was intrigued by this suggestion. After some discussion, it was agreed that clustering analysis should be done, and that a useful segmentation would be one in which for at least 2 of the 3 groups: a) the proportion of students with *NoPaymentDue* = ‘neg’ was very different from the corresponding overall proportion for the entire dataset; and b) the difference in the proportions of students with *NoPaymentDue* = ‘neg’ for the 2 groups was statistically significant.

**PART B**

##### B1. Introduction

For this part of the project you will use the following dataset:

Model Dataset: DMABASE.CSV

Target Variable: LOGSALAR

Below is the variable label and description.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Role** | **Measurement** | **Type** | **Variable Label** |
| NAME | id | nominal | char | Player's Name |
| TEAM | input | nominal | char | Team at the end of 1986 |
| NO\_ATBAT | input | interval | num | Times at Bat in 1986 |
| NO\_HITS | input | interval | num | Hits in 1986 |
| NO\_HOME | input | interval | num | Home Runs in 1986 |
| NO\_RUNS | input | interval | num | Runs in 1986 |
| NO\_RBI | target | interval | num | RBIs in 1986 |
| NO\_BB | input | interval | num | Walks in 1986 |
| YR\_MAJOR | input | interval | num | Years in the Major Leagues |
| CR\_ATBAT | input | interval | num | Career times at bat |
| CR\_HITS | input | interval | num | Career Hits |
| CR\_HOME | input | interval | num | Career Home Runs |
| CR\_RUNS | input | interval | num | Career Runs |
| CR\_RBI | input | interval | num | Career RBIs |
| CR\_BB | input | interval | num | Career Walks |
| LEAGUE | input | binary | char | League at the end of 1986 |
| DIVISION | input | binary | char | Division at the end of 1986 |
| POSITION | input | nominal | char | Position(s) in 1986 |
| NO\_OUTS | input | interval | num | Put Outs in 1986 |
| NO\_ASSTS | input | interval | num | Assists in 1986 |
| NO\_ERROR | input | interval | num | Errors in 1986 |
| SALARY | input | interval | num | 1987 Salary in $ Thousands |
| LOGSALAR | input | interval | num | Log Salary |

For each Task that is described in the next section you will have to do the following:

1. Develop and describe your plan for addressing the task that includes a detailed description of each step in the appropriate order. The description of your plan should not include material that you obtained from the execution of your plan! You must develop & document your plan before you execute it.
2. Execute your plan. Provide evidence of your work in a manner similar to that required for CE1, CE2, etc.

##### B2. Description of Tasks

**Task B1:**

Generate the ‘best’ Predictive Model using Average Square Error as the Model Assessment Measure. You should consider the use of derived variables.

**Task B2:**

Generate different clustering segmentations, where the maximum number of clusters should be less or equal to the number of rules in your choice of the most appropriate decision tree for Task B1. Provide the ‘best’ explanation for the clusters of this segmentation.

**Task B3:**

Partition the continuous variable LOGSALAR into three intervals: (0, Mean – 0.25\*StdDev), (Mean – 0.25\*StdDev, Mean + 0.25\*StdDev], (Mean + 0.25\*StdDev, +∝]. Given this partition, generate the best’ Decision Tree. You should consider the use of derived variables.