The Default of Credit Card Clients

Jesus Gomez and Jammy Loeur

Department of Computer Science
California State University
Sacramento, CA 95819, USA
jgomez42gb@gmail.com, jammyloeur09@gmail.com

Abstract—*Team nine members, Jesus Gomez and Jammy Loeur, utilized a dataset obtained from the UCI Machine Learning Repository website to gain knowledge in data warehousing and data mining applications.*

Through months of research and studies, Jesus and Jammy were able to successfully complete their data science projects and create an online data mart where users could perform analytical examination on their dataset. Additionally, they gained substantial knowledge with the Naive Bayesian classification method as well as the process of knowledge discovery through data mining.

1.0 DATA WAREHOUSE INTRODUCTION

The aim of our data mart project is to produce a functional system capable of retrieving records for query and analysis. In building our data mart, we utilized a dataset consisting of credit card records. We are interested in analyzing the different attributes that most lead to a person defaulting on their next credit card payment. Some of the questions we wanted to answer using our data mart are:

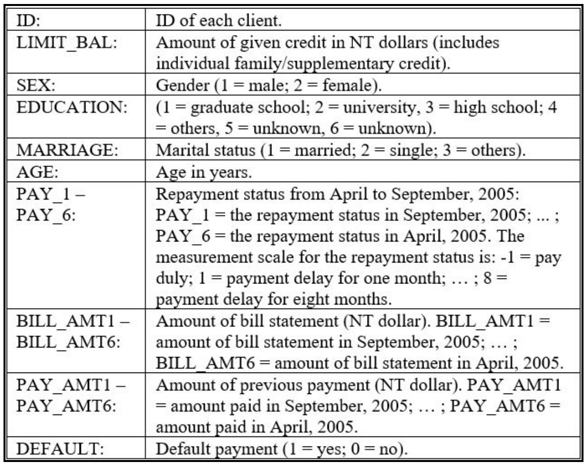
- ➢ What customers are defaulting on their payments?
- ➢ What is the proportion of payment defaults for specific customers?

For the construction of our data mart we used the star schema model to base our MySQL database design. We also transformed our data and finally, loaded it onto a MySQL database and developed a web interface using PHP to query from the data mart.

1.1 DATASET DESCRIPTION

The dataset chosen for this project was acquired from the University of California, Irvine Machine Learning Repository [1]. It consists of 30,000 records of Taiwanese credit card customers and with a total of 25 attributes relating to payment defaults, customer demographics, credit data, history of payments, and bill statements from April to September 2005. An overview of the dataset can be seen in Table 1.

Table 1. Description of our Dataset.



ID:	ID of each client.
LIMIT_BAL:	Amount of given credit in NT dollars (includes individual family/supplementary credit).
SEX:	Gender (1 = male; 2 = female).
EDUCATION:	(1 = graduate school; 2 = university, 3 = high school; 4 = others, 5 = unknown, 6 = unknown).
MARRIAGE:	Marital status (1 = married; 2 = single; 3 = others).
AGE:	Age in years.
PAY_1 – PAY_6:	Repayment status from April to September, 2005: PAY_1 = the repayment status in September, 2005; ... ; PAY_6 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; … ; 8 = payment delay for eight months.
BILL_AMT1 – BILL_AMT6:	Amount of bill statement (NT dollar). BILL_AMT1 = amount of bill statement in September, 2005; … ; BILL_AMT6 = amount of bill statement in April, 2005.
PAY_AMT1 – PAY_AMT6:	Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September, 2005; … ; PAY_AMT6 = amount paid in April, 2005.
DEFAULT:	Default payment (1 = yes; 0 = no).

1.2 DATA MART DESIGN

For our data mart design we decided to utilize the star schema to model it as it allows for simpler queries and better query performance. The star schema we created consists of one fact table referencing four dimension tables. The fact table includes foreign keys, that link to each dimension table, and the customer default status attribute. The dimension attributes of our star schema are comprised of the following: customer information, previous payment, bill statement, and repayment status. The subsequent star schema can be viewed in Figure 1 below.

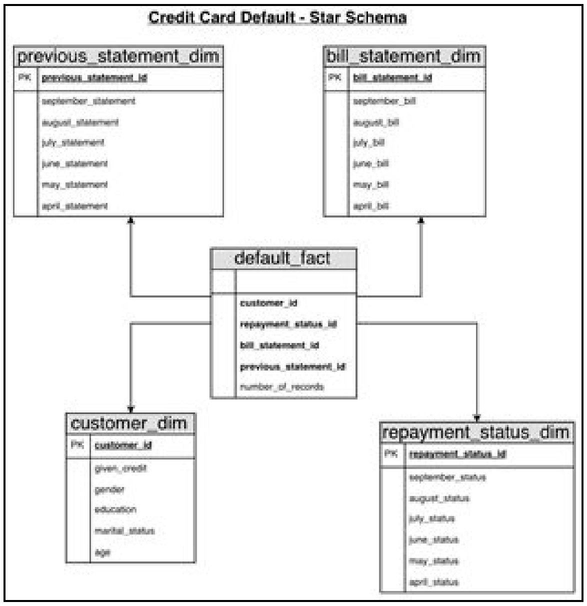


Figure 1. Schema Design used for Data Mart.

1.3 DATA PREPROCESSING

To accomplish the task of building our data mart, our team's first objective was to properly prepare the dataset using various preprocessing methods. Our dataset was in a fairly clean state when we first downloaded it from the Machine Learning Repository, but there were still some minor issues to be corrected.

1.3.1 DATA CLEANING

The issues we primarily found consisted of data values that were undefined by the UCI website; one example was the EDUCATION attribute in which values one through four were defined but it also contained values that were greater than four. In these instances, we used a tool named RStudio, a free and open source IDE for R [8], to re encode the undefined values into the more appropriate "Others" category that the attribute information specified. We repeated did this procedure to the MARITAL_STATUS attribute, as it contained the same error as our EDUCATION column.

1.3.2 DATA TRANSFORMATION

The next step in preprocessing was to transform the data. The dataset from the UCI Machine Learning Repository came completely optimized for data mining which meant all the record values were in numerical form; for example, categorical data such as SEX were encoded as "1" and "2" to represent male and female respectively. This was an issue as having numerical values in our data mart would diminish its effectiveness for our users. Hence, we needed to transform specific columns so that they we would be more suitable for use in the data mart. This involved using RStudio and Kabacoff's comprehensive guide in the use of R [10], to re encode the attributes SEX, EDUCATION, and MARITAL_STATUS into their string representations.

1.3.3 DATA SPLITTING

Our final preprocessing step was to divide the dataset into their proper fact and dimension tables specified by the star schema and then load the resulting tables onto a MySQL database. The process of splitting the dataset into separate tables was easily accomplished with RStudio which we could use to generate the five csv files needed for our database design. To convert our csv files to MySQL statements we used an online tool called ConvertCSV which greatly aided in our process of converting of our dataset files for our database [9].

1.4 IMPLEMENTATION

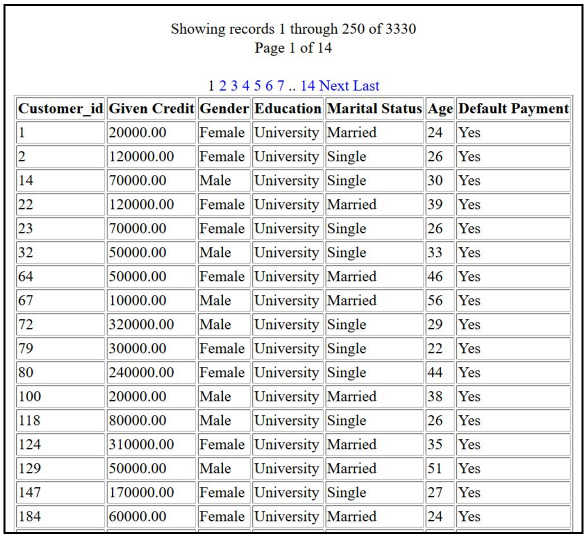
With our dataset now on a MySQL database, we proceeded to create a web page to access our database and query it using user specified parameters. The intended goal of our data mart was to build a simple, yet functional interface that would allow users to retrieve records from our dataset.

The web page that hosted the data mart was built using a combination of PHP, HTML, and CSS at the front end, while MySQL and the Athena Web Server were used as the back end of the system. We used many tutorials from W3Schools in the development of our web pages as well as the construction of the data mart [5]. We were able to create a working data mart fairly quickly, but we continued making improvements on it late into the semester.

1.5 RESULTS

The product of our work and studies of data warehousing led to the creation of a data mart that is effective in retrieving user defined queries and is also easy to use.

Users can retrieve customer information using various drop down menus that then get queried and then used by our database. The information that fits the user's criteria are retrieved and the first 250 results are displayed in a table format, as shown in Figure 2. We added the constraint of limiting the results per page to prevent performance issues that were caused when large amounts of data are retrieved. Additionally, our data mart displays record information and navigation buttons to transition between pages to provide the best user experience with our data mart.



Showing records 1 through 250 of 3330
Page 1 of 14

1 2 3 4 5 6 7 .. 14 Next Last

Customer_id	Given Credit	Gender	Education	Marital Status	Age	Default Payment
1	20000.00	Female	University	Married	24	Yes
2	120000.00	Female	University	Single	26	Yes
14	70000.00	Male	University	Single	30	Yes
22	120000.00	Female	University	Married	39	Yes
23	70000.00	Female	University	Single	26	Yes
32	50000.00	Male	University	Single	33	Yes
64	50000.00	Female	University	Married	46	Yes
67	10000.00	Male	University	Married	56	Yes
72	320000.00	Male	University	Single	29	Yes
79	30000.00	Female	University	Single	22	Yes
80	240000.00	Female	University	Single	44	Yes
100	20000.00	Male	University	Married	38	Yes
118	80000.00	Male	University	Single	26	Yes
124	310000.00	Female	University	Married	35	Yes
129	50000.00	Male	University	Married	51	Yes
147	170000.00	Female	University	Single	27	Yes
184	60000.00	Female	University	Married	24	Yes

Figure 2. Our Data Mart Results Page.

1.6 KNOWLEDGE GAINED, DW

Our team obtained a great deal of valuable learning experiences through the design and implementation of the data mart. Although designing the structure of the data mart was not particularly difficult, the real challenge came from the process of implementing it. We had never used RStudio before so there was a learning curve to effectively use this data analysis tool. The provided R Demos from the course website greatly aided in the process of learning this new tool, as well the numerous websites on this topic. The biggest learning experience came from the creation of our data mart as we had little to no prior experience working with web development languages and tools such as PHP, and MySQL retrieval from a web page. Overall, this portion of the project was very beneficial for both of us as we learned an abundant number of new things.

2.0 DATA MINING INTRODUCTION

In the first portion of our project we performed some minor data preprocessing on our dataset to make it more suitable for data mart application, but in order to conduct efficient and worthwhile data analysis on our dataset, we would need to use more intensive preprocessing methods and attribute examination to determine the best approach for our dataset. We experimented with various classification methods such as support vector machine, c5.0, decision tree classification and found Naive Bayesian to be the best fit for our situation. We used this classification method and several other resources to discover the meaningful patterns within our data and answer the following questions:

- ➢ What combination of customer attributes will maximize the probability of default?
- ➢ ... minimize the probability of default?
- ➢ What customer attribute is the most influential to the probability of payment default?

To accomplish the tasks of answering our questions and learning more about the dataset, we utilized several tools such as R and RStudio for both data preprocessing and data mining, Tableau for data visualization and analysis, Excel for general purpose use, and naive bayesian for our method of data mining.

2.1 RELATED WORK

Similar work was performed by a group of researchers who used the same dataset to measure the performance of six different classification methods. Unlike our project, the researchers were predicting the probability values of the dataset. Their findings concluded that logistic regression achieved the best performance when compared to other classification methods [2]. Their research studies served as an example of how we should conduct our data mining segment of the project with regards to both quality and procedure.

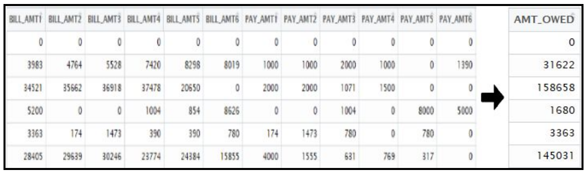
2.2 DATA PREPROCESSING

The classification method that we chose initially posed a problem with our dataset as this data mining technique required attributes that were independent of one another. Our preliminary application of the naive bayes gave us accuracy rates of 59-60%, too low for any practical use in our analysis. Thus, we needed to apply data preprocessing to improve the accuracy of our classification and gain useful results. To achieve a suitable dataset for naive bayesian we derived new and more useful attributes, implemented data discretization, and applied target class balancing as well to our data.

2.2.1 ATTRIBUTE CONSTRUCTION

The initial dataset was both highly correlated and high-dimensional. This greatly contributed to the inadequate performance of our classifier. To reduce the impact of this issue we applied attribute construction and replacement on the dataset. Using the values of BILL_AMT1 through BILL_AMT6 and PAY_AMT1 through PAY_AMT6, we derived a new attribute, AMT_OWED, by subtracting the total sum of BILL_AMTs from the total sum of PAY_AMTs for each record. This can be seen in Table 2.

Table 2. Attribute Construction



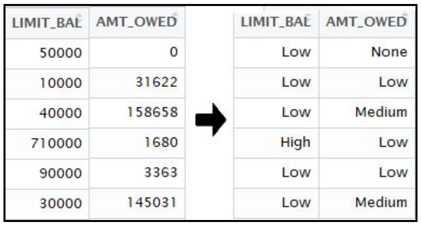
BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6		AMT_OWED
0	0	0	0	0	0	0	0	0	0	0	0		0
3983	4764	5528	7420	8298	8019	1000	1000	2000	1000	0	1390		31622
34521	35662	36918	37478	20650	0	2000	2000	1071	1500	0	0	➡	158658
5200	0	0	1004	854	8626	0	0	1004	0	8000	5000		1680
3363	174	1473	390	390	780	174	1473	780	0	780	0		3363
28405	29639	30246	23774	24384	15855	4000	1555	631	769	317	0		145031

We then removed the original attributes BILL_AMTs from the dataset as they were no longer needed and would affect the outcome of our classifier. Additionally, we derived another attribute, MISSED_PAYMENTS, from the values of PAY_1 through PAY_6, which we also removed from the dataset. In summation, we replaced 18 attributes with two derived ones which led to the outcome of uncorrelated attributes that still held the initial intent of the original attributes.

2.2.2 DATA DISCRETIZATION

The second preprocessing step that was needed was to apply data discretization, or the technique of converting a continuous variable into a discrete one by creating a set of intervals that stretches the range of the variable's values. We applied discretization to our dataset as a group of researchers were able to conclude that discretization improves the performance of the naive bayesian algorithm [3]. The attribute AGE was split into equal sized intervals of ten years and both attributes, AMT_OWED and LIMIT_BAL, were placed into buckets labeled as either "Low", "Medium", or "High" indicating the level of amount owed and credit limit accordingly. The method for discretizing the two attributes was chosen subjectively. We defined ranges, in New Taiwan dollars, of "$0-$100,000 NT", "$100,001-$500,000 NT", and over "$500,001 NT" to the labels respectively. The result of this can be seen in Table 3 below.

Table 3. Data Discretization



LIMIT_BAL	AMT_OWED
50000	0
10000	31622
40000	158658
710000	1680
90000	3363
30000	145031

➡

LIMIT_BAL	AMT_OWED
Low	None
Low	Low
Low	Medium
High	Low
Low	Low
Low	Medium

2.2.3 TARGET CLASS BALANCING

Our third and final preprocessing step that we applied was data reduction. In reference to a discussion on Data Science Stack Exchange, we decided to reduce the size of the dataset to achieve an equal class representation between the default and no default classes to create the predictive model for our data [7]. The dataset was reduced from 30,000 to 13,272 records, with a 50% class representation of 6,636 records each. Additionally, we removed any irrelevant and correlated data from our dataset in order to facilitate a more suitable dataset for naive bayes use. Our original dataset contained 25 attributes but after preprocessing was complete we had seven, two of which were derived from the original attributes.

2.3 IMPLEMENTATION

With data preprocessing complete, we could start the data mining process. The first step in any classification method is to train the classifier to correctly label the class attribute in our dataset. For this we applied the holdout procedure and partitioned our data into two sets, 80% training set and a 20% testing set. The result of this splitting created a training set containing 10,618 records and a testing set containing 2,654 records. Our team then applied the naive bayesian algorithm to the training set and built a classifier using RStudio. The implementation of the naive bayesian algorithm was relatively simple to accomplish due to the fact there were several resources available to guide us through the creation of the R scripts needed for this method. This classifier was then applied to the testing set for performance evaluation.

2.4 FINDINGS

The results of our preprocessing and careful implementation of naive bayesian led to the creation of several useful outputs for us to demonstrate our knowledge of data mining. We used RStudio's confusion matrix script to generate the confusion matrices for our training and testing sets which then were used to calculate the accuracy rate of our model as shown in the text, *Introduction to data mining* [6]. We found that our preprocessed data achieved an accuracy rate of 71%, an 11% increase from the unaltered dataset accuracy, as shown by Table 4.

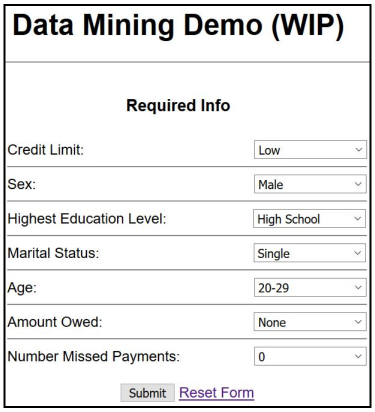
Table 4. Accuracy Rate Comparison.

	Training	Testing
Before Preprocessing	0.59	0.60
After Preprocessing	0.69	0.71

Furthermore, we were able to obtain the conditional probabilities of our training set and, using Example 8.4 in the text *Data mining: concepts and techniques*, use these rules to answer our initial questions regarding this portion of the project [4]. We learned that,

- ➢ Customers who are males, with a high school education, married, and are between 60 and 69 years old have the highest rate of payment default with a 53% likelihood.
- ➢ Customer with the least chance of defaulting are single females, who have a graduate education, between the ages of 30 and 39 have a 53% of paying on time.
- ➢ The attribute with the highest correlation with payment default was the age bracket of 60-69 with a 56% chance of defaulting.

We were also able to further use our conditional probabilities to create a classifier that uses several customer attributes to predict the chance of default as shown in Figure 3.



Data Mining Demo (WIP)

Required Info

Credit Limit:	Low
Sex:	Male
Highest Education Level:	High School
Marital Status:	Single
Age:	20-29
Amount Owed:	None
Number Missed Payments:	0

Figure 3. Data Mining Classifier Demo.

These attributes include customer demographics, credit and payment history. When users select the values they are interested in, our classifier will label the customer as either defaulting or not defaulting on their next credit card payment with an accuracy of 71%. It's a very simple application but very interesting nonetheless.

2.5 KNOWLEDGE GAINED, DM

The entire data mining process was a learning experience for both of us as we did not have any prior knowledge with this type of work coming in. By going through the individual steps of data mining ourselves, we were able to delve deeper into the concepts of data mining and its applications than would be possible through a textbook.

We gained first-hand experience with the process of knowledge discovery, preprocessing, data mining, evaluation, presentation. We discovered that preprocessing was the most important step in data mining. Our dataset required plenty of preprocessing and as such we became very familiar with this step by the end of the semester. We learned firsthand the importance of this initial step and the impact it can have on the data mining results through our own experience with it. Additionally, we obtained a more thorough understanding of the operation and limitations of the naive bayes classification method. Lastly, we learned how to operate data analysis and visualization tools such as RStudio and Tableau to evaluate our dataset, determine the accuracy of our methods, and to create useful histograms to illustrate the contents of our data to others.

3.0 CONCLUSION

The data science term project was a highly rewarding experience. We took an ordinary dataset and extracted meaningful information that could be adopted to real world applications. Through each stage of the project we learned a host of new materials as well as skills; from the design and construction of a data mart that allows users to query records, to the data preprocessing and the application of naive bayes to gain knowledge from data. Our time working on this data science term project has opened an exciting avenue in computer science that our team has not yet encountered but perhaps may further explore in the future.

4.0 REFERENCES

[1] UCI Machine Learning Repository: default of credit card clients Data Set. Accessed March 16, 2017. http://archive.ics.uci.edu/ml/datasets/default+ of+credit+card+clients.

[2] Yeh, I-Cheng, and Che-Hui Lien. "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients." *Expert Systems with Applications* 36, no. 2 (2009): 2473-480. doi:10.1016/j.eswa.2007.12.020.

[3] Lustgarten JL, Gopalakrishnan V, Grover H, Visweswaran S. Improving Classification Performance with Discretization on Biomedical Datasets. *AMIA Annual Symposium Proceedings*. 2008;2008:445-449.

[4] Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Amsterdam: Elsevier/Morgan Kaufmann, 2012.

[5] PHP 5 Tutorial. Accessed April 24, 2017. https://www.w3schools.com/php/default.asp.

[6] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Dorling Kindersley: Pearson, 2015.

[7] "Should I go for a 'balanced' dataset or a 'representative' dataset?" Machine learning - Should I go for a 'balanced' dataset or a 'representative' dataset? - Data Science Stack Exchange. Accessed April 29, 2017. https://datascience.stackexchange.com/questions/810/should-i-go-for-a-balanced-dataset-or-a-representative-dataset/8628#8628.

[8] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessed April 12, 2017. https://www.R-project.org/.

[9] “Convertcsv.com - Convert CSV to JSON, XML, SQL,.” Accessed April 26, 2017. http://www.Convertcsv.com.

[10] Kabacoff, Robert. "R Tutorial." R Tutorial For Beginners. Accessed April 9, 2017. http://www.statmethods.net/r-tutorial/index.html.