Credit Card Default

Team 09
Jesus Gomez and Jammy Loeur

Outline of our Presentation

I. INTRODUCTION
   i) Our Dataset

II. DATA WAREHOUSING
   i) Objective
   ii) Data Preprocessing
   iii) Implementation
   iv) Data Mart Demo

III. DATA MINING
   i) Motivation
   ii) Representative Models
   iii) Data Preprocessing
   iv) Predictive Models
   v) Findings

IV. CONCLUSION
   i) Knowledge Gained
   ii) Challenges Faced
   iii) References
   iv) Questions

Introduction

Everything you need to know about our dataset!

Our Dataset

- Dataset contains customer information and their payment history over a six month period.
  - Time period, April - September 2015
  - Target class is the default status of the month of November 2015
  - Dataset information originates from studies conducted in Taiwan

World Map

30,000 Records

That's a lot of records!

24 Attributes

Education, sex, age, pay status, bill and pay amounts, default status!

Source:

https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Question

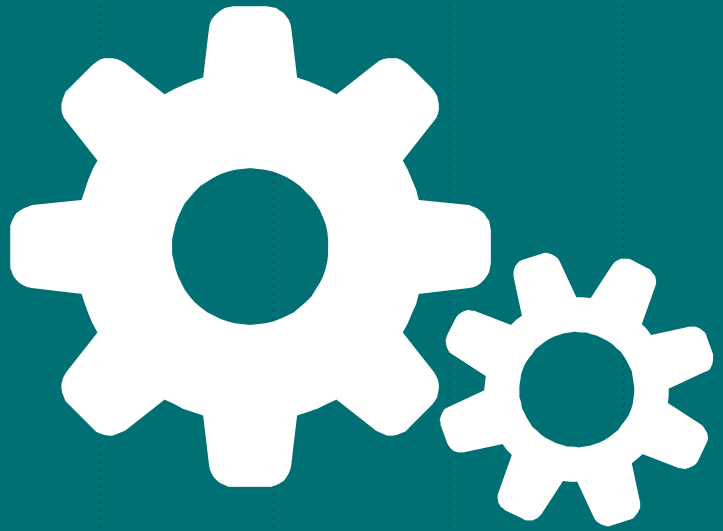
What does our dataset deal with?

A. Gift Cards
B. Credit Cards
C. Trading Cards
D. Birthday Cards

Question

What does our dataset deal with?

B. Credit Cards



Data Warehousing

Creating a data mart for analytical purposes!

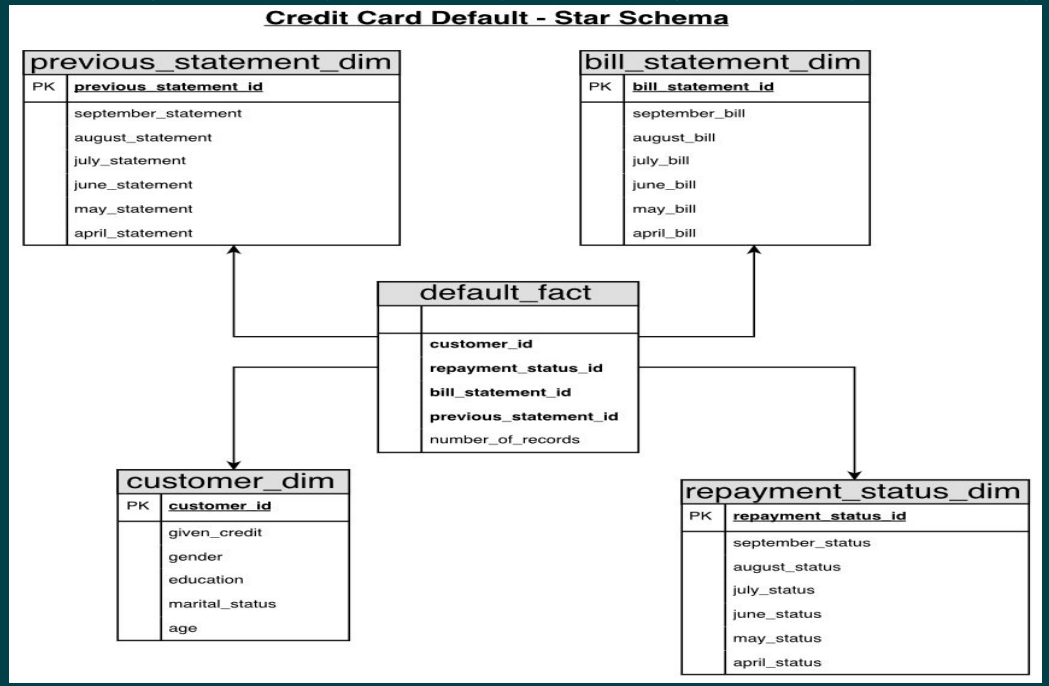
Objective

- Create a functional data mart capable of retrieving records for our users

Motivation

- What customers are defaulting on their payment?
- What is the proportion of payment defaults with certain customer attributes?

Star Schema


Credit Card Default - Star Schema
previous_statement_dim
PK
previous_statement_id
september_statement
august_statement
july_statement
june_statement
may_statement
april_statement
bill_statement_dim
PK
bill_statement_id
september_bill
august_bill
july_bill
june_bill
may_bill
april_bill
default_fact
customer_id
repayment_status_id
bill_statement_id
previous_statement_id
number_of_records
customer_dim
PK
customer_id
given_credit
gender
education
marital_status
age
repayment_status_dim
PK
repayment_status_id
september_status
august_status
july_status
june_status
may_status
april_status

Question

What schema design did we implement?

A. Snowflake Schema
B. Raindrop Schema
C. Star Schema
D. Fact Constellation

Question

What schema design did we implement?

C. Star Schema

Implementation

- Database
  - ConvertCSV
  - MySQL
- Data Mart
  - Front-End
    - PHP
    - HTML / CSS

Data Preprocessing

- Cleaning
  - Dataset contained undefined values.
  - Reencode values into proper category.
- Transformation
  - Dataset completely in numeric form.
  - Convert Marital Status, Sex, and Education attributes into strings for readability.
    - 1 -> “Male”
    - 2 -> “Female”

Adding the Data to MySQL

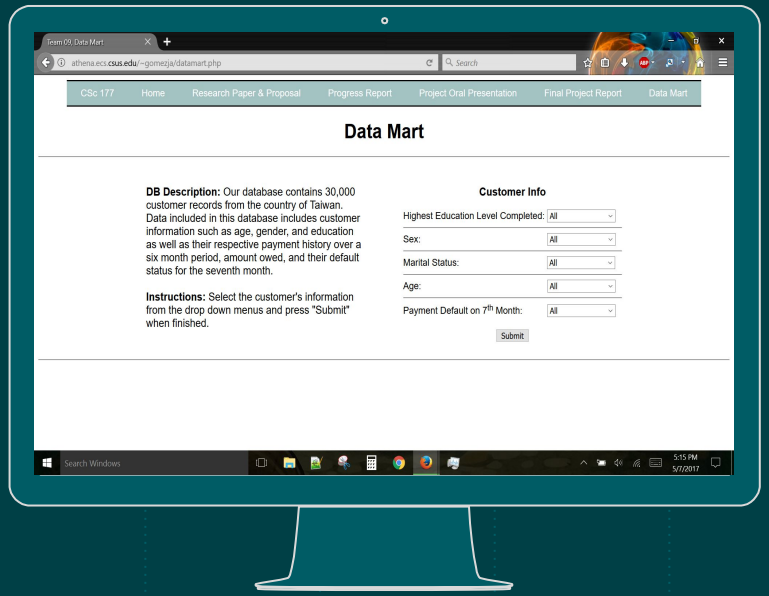
- Create Customer dimension

```
CREATE TABLE customer_dim (
        customer_id int NOT NULL,
        given_credit decimal(10,2),
        gender varchar(255),
        education varchar(255),
        marital_status varchar(255),
        age int,
        PRIMARY KEY (customer_id)
);
```

- Insert data into Customer table

```
/* INSERT QUERY */INSERT INTO customer_dim(customer_id,given_credit,gender,education,marital_status,age) VALUES(
1,20000,'Female','University','Married',24);
/* INSERT QUERY */INSERT INTO customer_dim(customer_id,given_credit,gender,education,marital_status,age) VALUES(
2,120000,'Female','University','Single',26);
/* INSERT QUERY */INSERT INTO customer_dim(customer_id,given_credit,gender,education,marital_status,age) VALUES(
3,90000,'Female','University','Single',34);
/* INSERT QUERY */INSERT INTO customer_dim(customer_id,given_credit,gender,education,marital_status,age) VALUES(
4,50000,'Female','University','Married',37);
/* INSERT QUERY */INSERT INTO customer_dim(customer_id,given_credit,gender,education,marital_status,age) VALUES(
5,50000,'Male','University','Married',57);
```

Data Mart



http://athena.ecs.csus.edu/~gomezja/datamart.php

Data Mining

Gaining knowledge from data!

Motivation

- To learn…
  - What combination of customer characteristics will maximize the probability of payment default.
    - Minimize the probability?
  - What customer attribute is the most influential to the probability of payment default.

Tools / Software Used

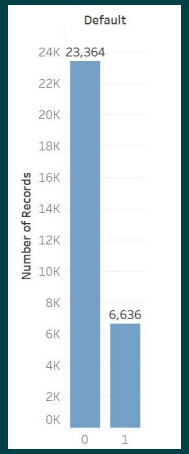
- R / RStudio
  - https://www.rstudio.com/
- Tableau
  - http://www.tableau.com/public/
- Excel
  - https://products.office.com/en-us/excel

Algorithm

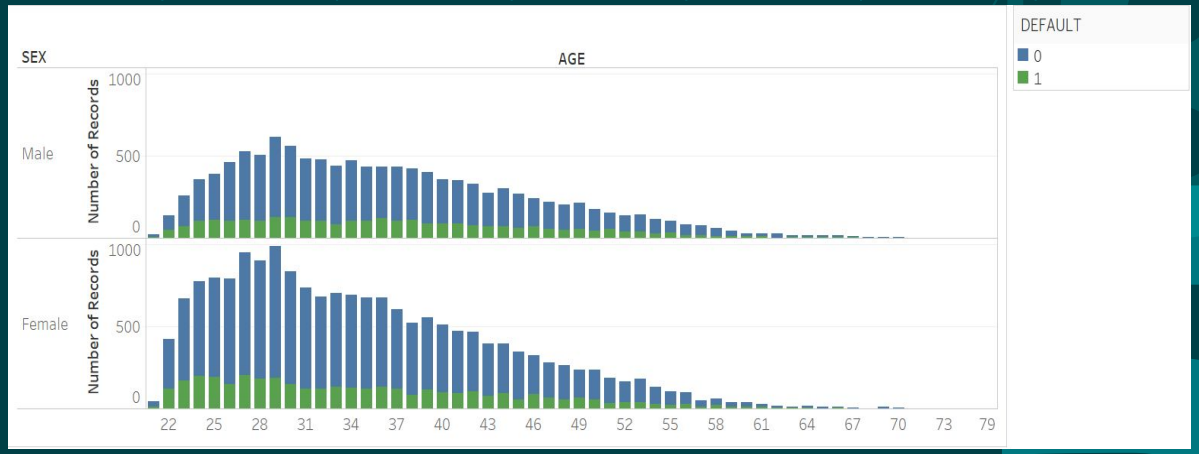
- Naive Bayesian Classification

Data Exploration

- Our data BEFORE preprocessing



Target class distribution



Relation between SEX, AGE, and DEFAULT

Data Preprocessing

- Data Reduction
  - Numerosity Reduction
    - Class imbalance in our data
    - Reduced dataset from 30,000 to 13,272
  - Dimensionality Reduction
    - Removed customer ID column
- Data Transformation
  - Attribute construction
  - Discretization

Data Preprocessing Cont.

- Discretization
  - Raw values of numeric attributes are replaced with interval labels or conceptual labels

```
dataset$SEX <- factor(dataset$SEX,
                      levels = c(1,2),
                      labels = c("Male", "Female"))
dataset$EDUCATION <- factor(dataset$EDUCATION,
                      levels = c(1,2,3,4),
                      labels = c("Graduate School", "University",
                            "High School", "Others"))
dataset$MARRIAGE <- factor(dataset$MARRIAGE,
                      levels = c(1,2,3),
                      labels = c("Married", "Single", "Others"))
```

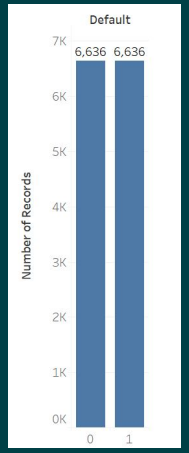
Conceptual labels

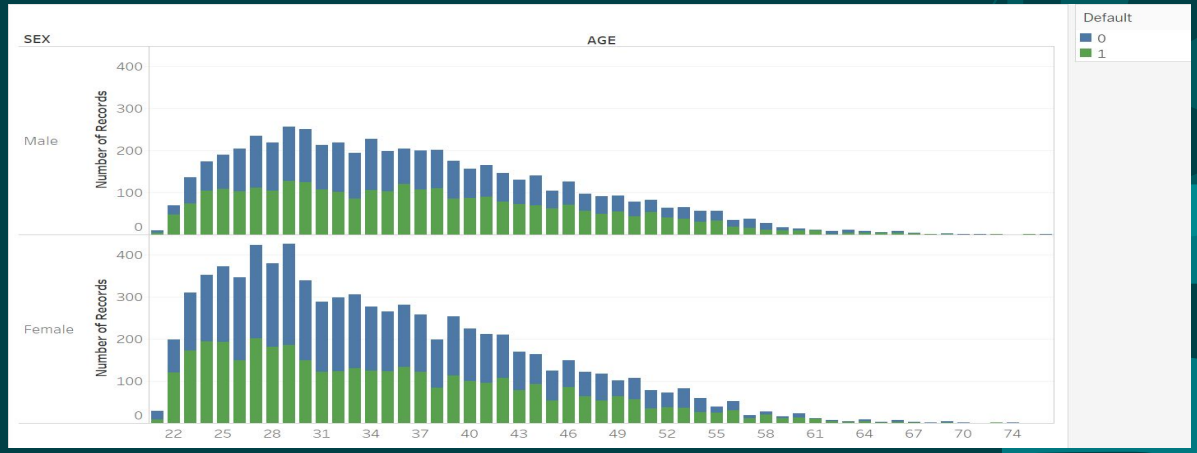
```
dataset$AGE = cut(dataset$AGE,
                  breaks = c(-Inf, 30, 40, 50, 60, 70, Inf),
                  labels = c("20-29", "30-39", "40-49", "50-59",
                             "60-69", "70-80"),
                  right = FALSE)
dataset$LIMIT_BAL = cut(dataset$LIMIT_BAL,
                        breaks = c(-Inf, 100000, 500000, Inf),
                        labels = c("Low", "Medium", "High"),
                        right = FALSE)
dataset$AMT_OWED = cut(dataset$AMT_OWED,
                       breaks = c(-Inf, 0, 100000, 500000, Inf),
                       labels = c("None", "Low", "Medium", "High"),
                       right = FALSE)
```

Interval labels

Data Exploration

- Our data AFTER preprocessing





Target class distribution

Relation between SEX, AGE, and DEFAULT

Data Exploration Cont.

- Our data AFTER preprocessing

30,000 Records

24 Attributes

13,272 Records

7 Attributes

Question

How many attributes does our dataset have after data preprocessing?

A. 9
B. 24
C. 13,272
D. 7

Question

How many attributes does our dataset have after data preprocessing?

D. 7

Data Partition

- Holdout Procedure
- Training Set, 80%
  - 10,618 records
- Testing Set, 20%
  - 2,654 records

```
#Script to partition dataset

set.seed(123)
split = sample.split(dataset$DEFAULT, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
```

Confusion Matrix

Training Set - Accuracy: 69%

	1	0
1	3,255	2,054
0	1,190	4,119

Testing Set - Accuracy: 71%

	1	0
1	832	495
0	283	1,044

Accuracy Rate Comparison

	Training	Testing
Before preprocessing	0.59	0.60
After preprocessing	0.69	0.71

Conditional Probabilities

P(DEFAULT=0) = .500
P(DEFAULT=1) = .500

P(LIMIT_BAL=”Low” | DEFAULT=0) = .348
P(LIMIT_BAL=”Low” | DEFAULT=1) = .515
P(LIMIT_BAL=”Medium” | DEFAULT=0) = .617
P(LIMIT_BAL=”Medium” | DEFAULT=1) = .469
P(LIMIT_BAL=”High” | DEFAULT=0) = .035
P(LIMIT_BAL=”High” | DEFAULT=1) = .015

P(SEX=”MALE” | DEFAULT=0) = .389
P(SEX=”MALE” | DEFAULT=1) = .430
P(SEX=”FEMALE” | DEFAULT=0) = .611
P(SEX=”FEMALE” | DEFAULT=1) = .570

P(EDUCATION=”Graduate School” | DEFAULT=0) = .360
P(EDUCATION=”Graduate School” | DEFAULT=1) = .305
P(EDUCATION=”University” | DEFAULT=0) = .461
P(EDUCATION=”University” | DEFAULT=1) = .505
P(EDUCATION=”High School” | DEFAULT=0) = .163
P(EDUCATION=”High School” | DEFAULT=1) = .185
P(EDUCATION=”Others” | DEFAULT=0) = .018
P(EDUCATION=”Others” | DEFAULT=1) = .005

P(EDUCATION=”Graduate School” | DEFAULT=0) = .360
P(EDUCATION=”Graduate School” | DEFAULT=1) = .305
P(EDUCATION=”University” | DEFAULT=0) = .461
P(EDUCATION=”University” | DEFAULT=1) = .505
P(EDUCATION=”High School” | DEFAULT=0) = .163
P(EDUCATION=”High School” | DEFAULT=1) = .185
P(EDUCATION=”Others” | DEFAULT=0) = .018
P(EDUCATION=”Others” | DEFAULT=1) = .005

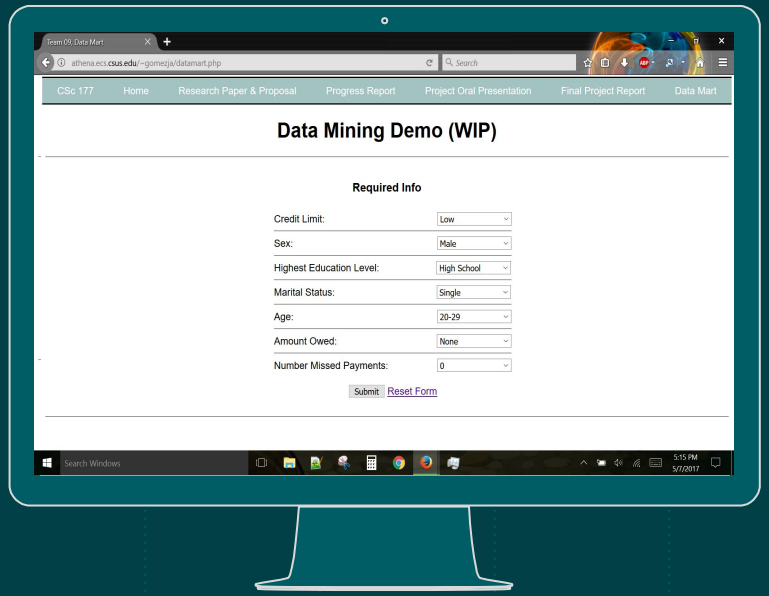
P(MARRIAGE=”Married” | DEFAULT=0) = 0.440
P(MARRIAGE=”Married” | DEFAULT=1) = 0.482
P(MARRIAGE=”Single” | DEFAULT=0) = 0.547
P(MARRIAGE=”Single” | DEFAULT=1) = 0.504
P(MARRIAGE=”Others” | DEFAULT=0) = 0.013
P(MARRIAGE=”Others” | DEFAULT=1) = 0.014

P(AGE=”20-29” | DEFAULT=0) = 0.327
P(AGE=”20-29” | DEFAULT=1) = 0.331
P(AGE=”30-39” | DEFAULT=0) = 0.382
P(AGE=”30-39” | DEFAULT=1) = 0.341

Findings

- What combination of customer characteristics will maximize the probability of payment default?
  X=(Sex="Male", Education="High School", Marital="Married", Age="60-69")
  - 53% chance of payment default
- Minimize the probability of payment default?
  Y=(Sex="Female", Education="Graduate School", Marital="Single", Age="30-39")
  - 53% chance of not defaulting
- What customer attribute is the most influential to the probability of payment default?
  - AGE="60-69"
    - 56% chance of payment default

Data Mining



http://athena.ecs.csus.edu/~gomezja/datamining.php

Conclusion

What we learned from this project!

Challenges Faced

- Finding the right classification algorithm
  - Low accuracy results
  - More data preprocessing was needed
- Data Preprocessing
  - Removing correlation between attributes values
  - Attributes were heavily correlated with one another
- Understanding the data set
  - Some attribute values were undefined
  - Very little documentation

Knowledge Gained

- Hands-on experience with the process of knowledge discovery
  - Preprocessing, data mining, evaluation, presentation
- Experience with Data Analysis & Visualization Tools
  - RStudio, Tableau
- Preprocessing is the most important step
  - Crucial part for proper data analysis
- Deeper understanding of classification
  - Algorithms and performance evaluation

References

- Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Waltham, MA: Morgan Kaufmann, 2012. Print.
- 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* 36.2 (2009): 2473-480. Web.
- Dataset:
  - http://archive.ics.uci.edu/ml/
- Conditional Probabilities:
  - http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-8.html
- This Presentation Template:
  - http://www.slidescarnival.com/

Thanks!

Any questions?