Project

# Customer

Marquette University

Mr. Thomas Kaczmarek (Director, Master of Science: Computing)

## Business

The Basic business of my project is doing prediction on the future enrollment situation of our school. There are three main goals for my project:

1. According to the model built from the original data to do classification on the future student (Classification problem).
2. Predict how many students will be in the future (Time series Analysis problem).
3. Do Clustering for application records of applicants, to find predictive variables and issues in the business

## Contacts

Client: [thomas.kaczmarek@mu.edu](mailto:thomas.kaczmarek@mu.edu)

Project team: [duo.zhang@marquette.edu](mailto:duo.zhang@marquette.edu)

# Schedule

Generally, my project is an 11 weeks length Business Analytics project, my plan is to arrange the project into five parts. They are:

1. Business understanding process (Week 1).
2. Data explore and understanding (Week 2-4).
3. Data preprocessing. (Week 5-9).
4. Choose & Use DM algorithms (Week 10-11).
5. Conclusion (Week 11).

List all meetings with the customer and important milestones that were accomplished.

The meeting schedule is shown in Figure 1:

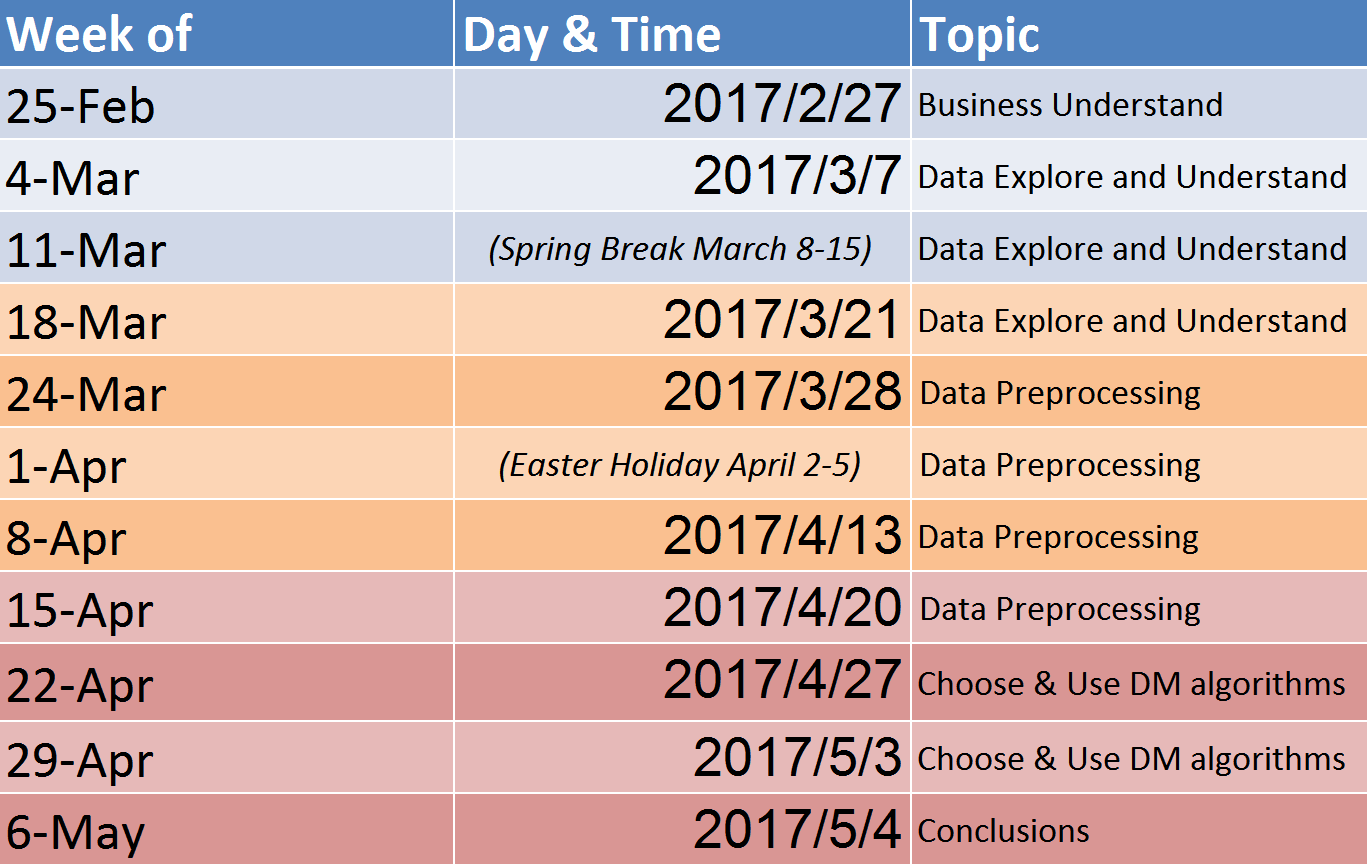
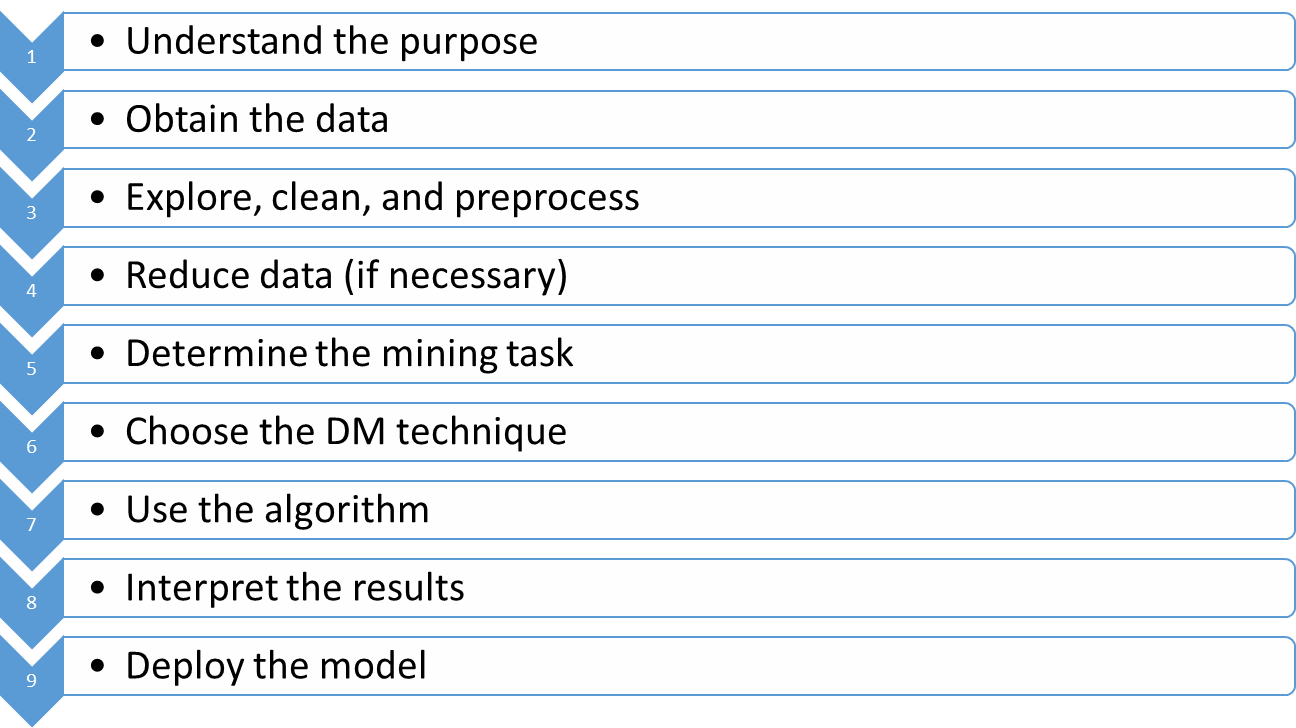


Figure 1

# Status

**Basic steps:**

1. understand the purpose: through talking with the client, I fully understand the purpose of this project, include what I need do, and what kind of result we should reach.
2. Obtain the data: I got the data in the second week from Mr. Thomas Kaczmarek.
3. Explore, clean, and preprocess: this is a big part of our project. Since the records are coming from the different time period of the whole apply process. Therefore, there are multiple records for each person. Our goal can be divided into three steps. First, split the data according to student ID. Second, merge all the records for a student for some specific rules. Third, combine all the new records into a new table.
4. Reduce data: according to our analysis we eliminate variables which have errors, and useless.
5. Determine the mining task: I choose four algorithms to do classification: K-N-N, Naïve Bayes, and (Classification tree), and Logistic Regression. Clustering I choose to use K-means.
6. Use the algorithm: I use them separately to build the model, and use clustering to choose predictive variables to rebuild the model.
7. Interpret the results: already did.
8. Deploy the model: This step, we need to consider according to different situations.

**What I learned from this project:**

1. The enrolling business of our school, and what are we looking for a student, when we judge whether we will send him/her an offer.
2. Using excel to deal with the data.
3. Visual Basic programming based on Excel.
4. Get a better understanding of each algorithm. Eg: When to deploy which algorithm, and what are the characteristics of them.
5. The basic steps of doing a Business Analytics project.
6. What are the important things should be paid more attention to, when one doing Business Analytics project.

## Purpose

The purpose of our project is doing prediction on the future enrollment situation of our school including: building a model to do classification on the future student, predicting how many students will be in the future, and doing clustering analysis.

## Data

What data did you acquire and what were the challenges. When? What did you learn in arriving at this?

Data I got:

1. Data of getting the data: Mar/7/2017.
2. Format: “.xlsx” excel file.
3. Quantity: 7036 records from 338 different students, for each student we got 31 variables.
4. History of the data: the recorded time for these records were from 2011-2017

Challenges:

1. Since the records for each student were from the different period of his or her application step, so it is hard to make a rule to combine them into a single record and do further mining.
2. After we finish making the merging rule, it is hard to make a Visual Basic Macro do filtering.

## Exploration

What did you do to explore the data before using it? When? What did you learn in arriving at this?

First Through the meeting, I understand all the abbreviations’ meaning. Second, I add a column called “the last status of the student” to see what the last status for each student, because the last status for a student is often more representative. Then, dealing with some error data and missing data. Next, we discuss the pattern of the records and get to know the business logic of this project.

Tasks to explore the data:

1. Understand the meaning of all the abbreviations (Project Meeting 2 Mar/7/2017).
2. Understand each category of the output variable, and start to recognize the pattern of it (Project Meeting 3-5).
3. Understand the language exam part. What score is expected for ESL students, and why the Quant is useful (Project Meeting 4-5).
4. Discuss the combination rules for the records of each student (Project Meeting 6).
5. Find solutions to split the data, and learning how to create a visual basic macro (Project Meeting 7-8).
6. Find solutions to combine multiple records of an applicant into a single one, still learning how to create a visual basic macro. (Project Meeting 9).
7. Improve the combination rules, and fix some errors when we do merging. (Project Meeting 10).

What did I learn:

Through all these processes, I learned several useful functions of excel using to deal with the data. Also, I learned how to create a Visual Basic macro based on excel. In addition, I know the basic to explore the data and preprocess the data.

## Data reduction if necessary

Did you eliminate any predictors/variables? Did you use any algorithmic technique? When? What did you learn in arriving at this?

Variables that have been eliminated:

1. Student ID
2. US Citizenship
3. State
4. Every variable that has relationship with GRE score except GRE date
5. Every variable that has relationship with TOEFL score except TOEFL TOTAL

I didn’t use algorithms for this part because we do not have numeric variables to deal with, and most of the operations are simple deleting. Through the reduction part, I get a better understand the business of our project.

## Task

What analysis approach or approaches did you decide to employ? When? What did you learn in arriving at this?

Approaches I decided to employ:

1. K-Nearest-Neighbor (classification) Apr/29/2017
2. Naïve Bayes (classification) Apr/29/2017
3. CART-Classification Tree (classification) Apr/30/2017
4. Clustering using all variables (find predictive variables) May/2/2017
5. Clustering using selected variables (find predictive variables) May/3/2017
6. Redo classification using K-Nearest-Neighbor (classification) May/4/2017
7. Redo classification using Naïve Bayes (classification) May/4/2017
8. Redo classification using CART-Classification Tree (classification) May/4/2017
9. Redo classification using Logistic Regression (classification) May/4/2017

What I have learned from this approach:

After I employ all the algorithms that I selected, I get a better understanding of each algorithm.

## Technique

What algorithms did you use? When? What did you learn in arriving at this?

The algorithm that has been used:

Classification:

1. K-Nearest-Neighbor (Apr/29/2017 – May/2/2017)
2. Logistic Regression (Apr/29/2017 – May/2/2017)
3. CART (Apr/29/2017 – May/2/2017)
4. Naïve Bayes (Apr/29/2017 – May/2/2017)

Clustering:

1. K-means (May/4/2017)
2. Hierarchical Clustering (May/4/2017)

Through this step, I get a better understand of each algorithm. I know how to tune each algorithm, and what reason may cause this problem.

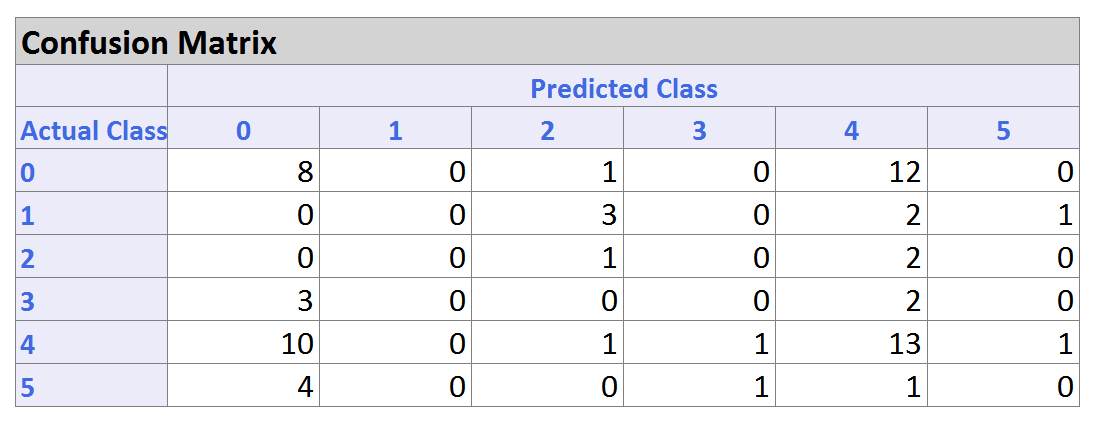
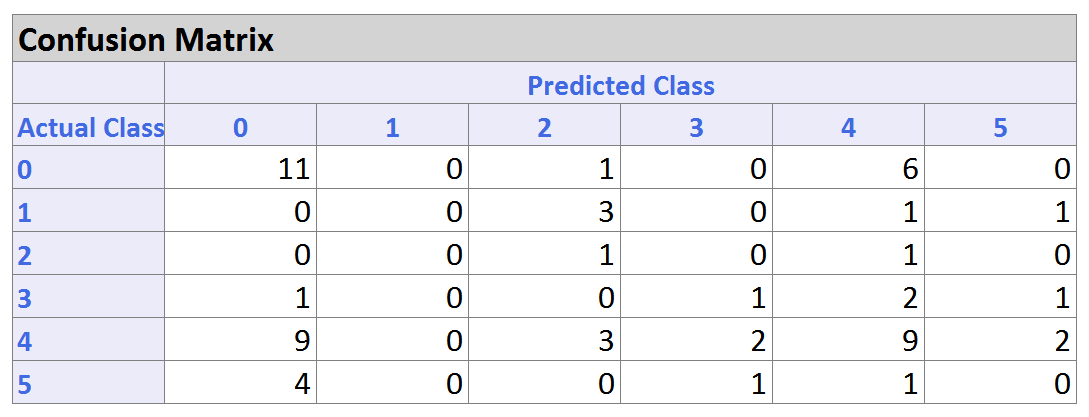
## Algorithm Result

What was the result of applying the technique(s)? When? What did you learn in arriving at this?

For this part, I will place the confusion matrix for test, and compare the error rate for each part。

**First classification result. (Apr/29/2017)**

K-N-N Naïve Bayes

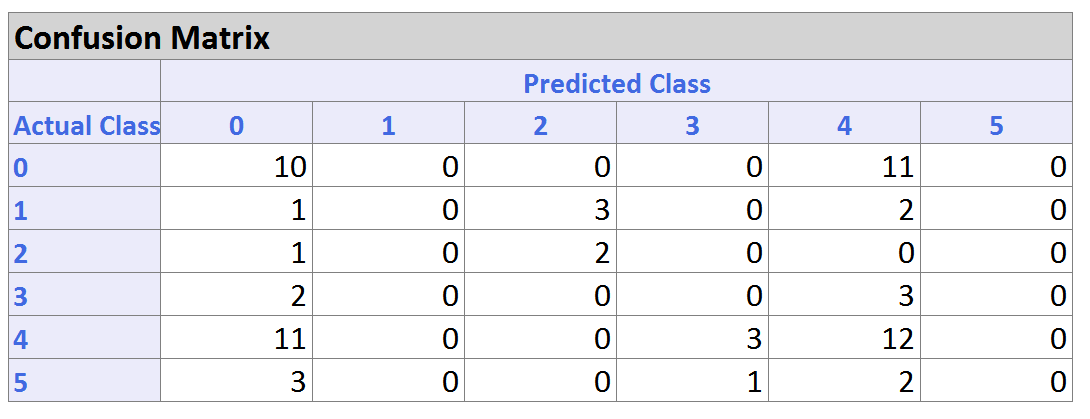
Error rate Error rate

Training：32.93% Training：38.92%

Validation：52% Validation：60%

Test：67% Test：63.93%

CART (classification tree)



Error rate

Training：0.60%

Validation：56%

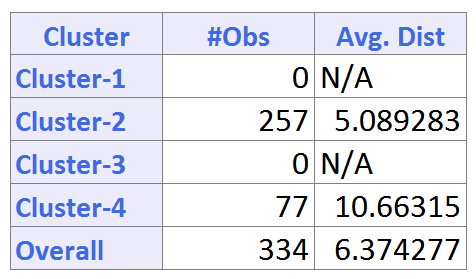
Test：64%

After the first classification, we found that the error rate for all algorithms is all bigger than 50%. Therefore, I plan to use clustering to find the more predictive variables.

**Clustering 1 (K-means)** **May/2/2017**

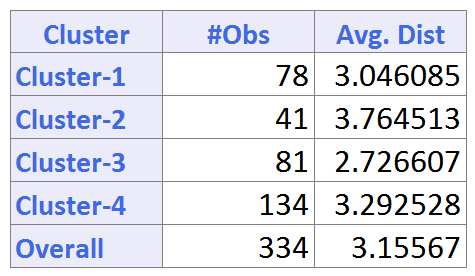
First, select all the predictors to do clustering

Select all the predictors, and do clustering (Do four clusters).



As we can see from the first clustering, there are two clusters have no objects. After analysis, I found that language score became a crucial classifier when we do clustering. Therefore, in the second clustering, I will delete all the variables that have relationships with language score.

**Clustering 2 (K-means)** **May/2/2017**



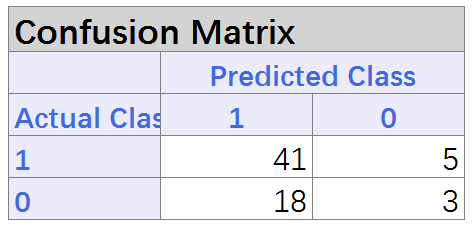
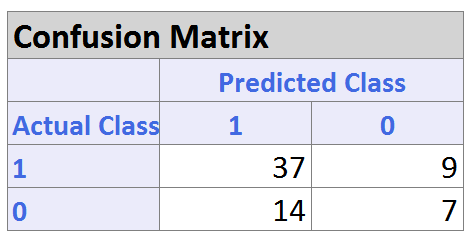
After I did the second clustering, through analysis I found 8 predictive variables, they are:

1. Sex
2. Financial Aid
3. Country \* 6

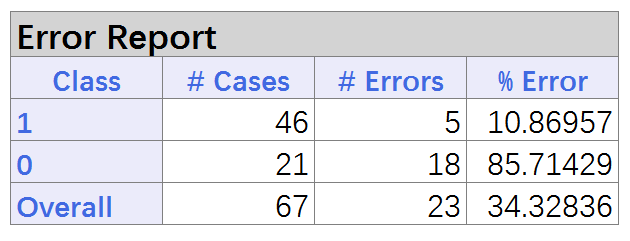
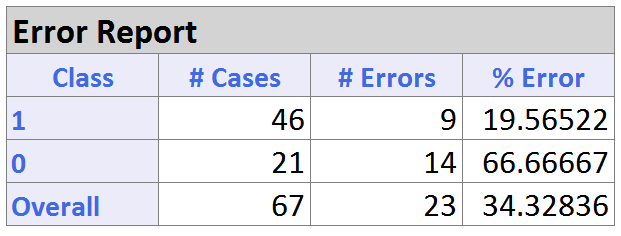
Using these predictive variables, I will do the second classification.

**Second classification result. (May/4/2017)**

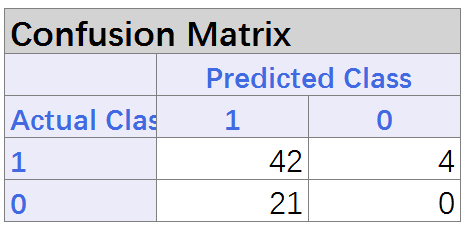
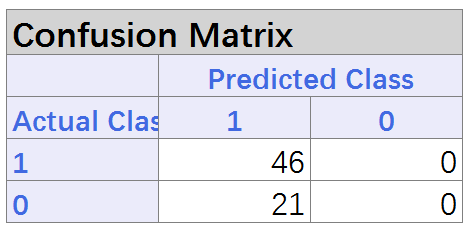
K-N-N Naïve Bayes

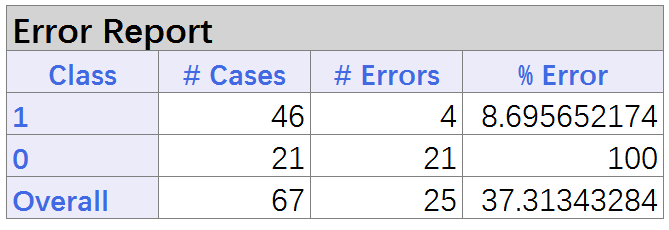
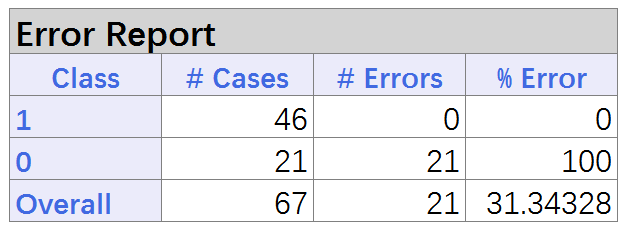
Error rate Error rate

Logistic Regression CART (classification tree)

Error rate Error rate

## Interpretation

According to Mr. Thomas Kaczmarek’s explanation, for the last classification result, we can see the error rate for class 1 is very low. In our project class 1 represent the student who will not matriculation. Therefore, my customer’s interpretation for this problem is: “According to the result of the last classification, our model has a marked ability to identify the student who will not become the student of Marquette University” (The last meeting May/4/2017)

Through this meeting, I learned:

When we are analyzing the result for algorithms, we can not only look at the overall error rate. We have to look at the error rate for each class, and the result of them will make more sense.

## Deployment/Consequence

Personally, I think the result will be used in the future work, may be in next academic year.

# Individual Contributions

Since the project is all finished by me, I have already explained the details in each part of the report. Therefore, I will not explain more in this part.

# Customer Benefits

1. Getting the excel macros to do processing and reduction on their data.
2. Getting the predictive variables (What information of a student is more important when they trying to judge whether we will send him/her an offer).
3. Getting the model to identify the student who will not matriculate (high accuracy).
4. Helping customer to find the issues in their business (After we did clustering, we find that student with some special characteristics have higher admission rate than other, and it may help our school to analysis what’s the factor lead this phenomenon).
5. Helping the customer to make a more effective plan. (According to our analysis, for some cities and countries there are more applicants than others, in the future, they can make a plan by focusing more and these areas on improving the efficiency.)

## Value to customer

Through this project, my customers get a detailed plan to process in the future data. In addition, they have also got the model to identify the students who will not matriculate.

## Customer satisfaction

On a scale from 1 to 10, how satisfied was the customer?

Very satisfied (10)

Customer suggestions.

No suggestion