PAKDD 2012 DATA MINING COMPETITION

http://fit.mmu.edu.my/pakdd2012/dmcomp.html

29th May-1st June 2012

Overview

The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) is pleased to organize a data mining competition. The competition is divided into two categories:

1. Open category (open to both academia, industry and students).
2. Student category (open to only student enrolled at institutions of higher learning). The prizes for the student category is sponsored by SAS Malaysia

The description of each category is available via the links below:

1. Go to the Open category
2. Go to the Student category link

PAKDD 2012 DATA MINING COMPETITION (STUDENT CATEGORY)

OVERVIEW

The theme of the competition centers on the discovery of correlation between the number of job applications and Malaysia's economic indicators.

This competition is open to only students of institutions of higher learning. The competition website can be accessed via PAKDD 2012 conference website (http://pakdd2012.pakdd.org/).

PROBLEM DEFINITION

There is only one problem to solve. The problem is described as follows:

Problem 1: Characterize the relationship(s) between the number of job applications and Malaysia's national economic indicators.

DATA SET

Only one data set is provided, and is available for download in the form of an Excel file. The data set consists of job application data and Malaysia's economic indicators from year 2006 to 2011. The Excel file has two tabs. The first tab titled "Job Application Data" has the following attributes:

- **Month:** This attribute gives the month of a year.
- **Total applications:** This gives the total job applications received through the data sponsor for a particular month.
- **Number of unique applicants:** This is the number of unique job applicants for a particular month.

The second tab in the Excel file is titled "Economic Indicators" and has the following attributes:

- **GDP:** This is the Gross Domestic Product at Purchasers' Value (measure as % of change). This is an indicator for the current economic growth in a particular quarter.
- **MARTCAP:** This is the market capitalisation of Bursa Malaysia (measured as annual change, %). Bursa Malaysia is Malaysia's Stock Exchange. This is an indicator for future economic growth.
- **CRECARD:** This is the outstanding balance of credit cards (measured as annual change, %).
- **CPI_A:** This is the Consumer Price Index (measured as % of change). The percentage of change is computed using the CPI of 2010 as a reference.
- **CPI_B:** The Consumer Price Index with the 2010 CPI as the reference.
- **EDUEXP:** This is the federal government's development expenditure in education.

- **FDI:** This is the amount of Foreign Direct Investment. The values for 2006 and 2007 are missing.

SUBMISSION

Competitors are required to submit the following through the competition website: A write-up (maximum 4 pages) describing the relationship(s) between the number of job applications and the economic indicators. The approaches and models that were used to discover the correlation must also be included in the write-up. This write-up must contain the names of the team members.

Participants are also asked to briefly describe how the SAS data mining tools were used to solve the problem.

DATA MINING TOOLS

Participants are highly encouraged to use the data mining tools from SAS (www.sas.com). The tools can be downloaded for trial

COMPETITION JUDGING

Submissions will be evaluated by a panel comprising of representatives from the organizer, the data sponsor and SAS Malaysia. The decision of the judging panel is final.

COMPETITION RULES

1. Students enrolled in any institution of higher learning who complies with the rules of the PAKDD 2012 Data Mining Competition may participate. The organizers and the data sponsor are not allowed to participate.
2. **Registration:**
   - All teams must register on the competition website in order to participate. To register a team, only the team leader needs to register.
   - Each student can only participate under one team. Multiple teams from the same university are allowed with the condition that the team leaders are different.
   - The names of team members must be disclosed when submitting the final entry. The composition of a team cannot be changed after the results are released.
3. Data are available for download only to registered participants.
4. **Challenge duration:** The challenge duration is about 4 weeks (16 April – 13 May 2012). To be eligible for prizes, competitors must submit their final entries by 13 May 2012. Each team is limited to one final entry. Only the last valid entry will be used for evaluation.
5. **Reproducibility:** Competitors do not have to deliver any code, but are required to produce a write-up (maximum 4 pages) to describe the results and the methods and models that they have used to solve the problems.
6. **Use of data:** The data may only be used for this competition.

PRIZES

Prizes will be allocated as follows:

Champion:	MYR 2000
1st Runner-Up:	MYR 1000
2nd Runner-Up:	MYR 500

Questions pertaining the competition should be directed to Lay Ki Soon at lksoon@mmu.edu.my

The prizes are sponsored by SAS Malaysia

OVERVIEW

- The theme of the competition centers on customer relationship management challenges faced by a large telco. The tasks for the competition focus on predicting customer churn and win-back.
- The competition is open to public, both industry and academic. The competition website can be accessed via PAKDD 2012 conference website (http://pakdd2012.pakdd.org/).

PROBLEM SUMMARY

- The PAKDD data mining competition 2012 provides two problems for participants to tackle. The tasks for each challenge are given below. Participants are requested to solve all tasks.
- The problems are motivated by challenges faced by a large telco. The problems involve two customer segments: Consumer (household users) and SME (Small and Medium Enterprises).
- The Consumer segment
- The consumer segment is concerned with voice and broadband services. Both services are considered as one integrated product for the purpose of this competition.
- The SME segment
- The SME segment is also concerned with voice and broadband services. However, they are treated as two separate products in the following way: voice and broadband as one product, and voice alone as another product by itself.

Maintaining customers and ensuring customers are satisfied with the products offered have always been a major challenge for telcos. Companies often try to reduce the number of customers terminating their services to avoid losing their market share. Before any intervention could be taken by any telco, a careful study of the characteristics of customers who churn must be done. Another important task is to win back customers who have switched to their competitors. In short, it is crucial for a company to know in advance the customers who have the intention to churn and identify the characteristics of those who could be won back.

Definitions:

1. Churn: A churn occurs when a customer terminates a particular service.
2. Winback: A winback occurs when a customer who has churned returns as an active subscriber of the service that he/she has previously terminated.

Problem 1 (Consumer Segment): Predicting potential churners and ex-customers who are likely to come back.

Tasks to Perform:

1. Predict churners: Produce a list of potential churners. Competitors are asked to submit the top 5% of existing customers who are most likely to churn.
2. Rank the drivers (i.e. the name of the attributes) that caused customers to churn.
3. Determine the top 5% of existing customers who are unlikely to churn, and why (i.e. provide the name of the attributes).
4. Predict win-back: Produce a list of ex-customers who are likely to come back as subscibers.

Problem 2 (SME Segment):

Tasks to perform:

1. Predict churners: Produce a list of potential churners. Competitors are asked to submit the top 5% of existing customers who are most likely to churn.
2. Rank the drivers (i.e. the name of the attributes) that caused customers to churn.
3. Determine the top 5% of existing customers who are unlikely to churn, and why (i.e provide the name of the attributes).

For Problem 2, competitors must submit their solutions for both products: voice and broadband, and voice alone. (See the explanation on the SME segment above).

DATA SET

Two data sets are provided, one for the Consumer segment, and another for the SME segment. Each data set consists of multiple tables. (For details, see the database design document.) The data sets consist of records of customer profiles (over more than 1 million customers), billing information, usage data, service requests and complaints. The data spans the period from 1 January 2011 to 31 December 2011. Within 2011, there were customers who churned, and there were win-back occurrences. The goal is to predict which customers (among those in the 2011 data) who will likely churn or return in the first three months of 2012. It is up to the contestants to divide the 2011 data into training and validation (hold-out) sets for their model development.

- A customer is uniquely identified using the attribute CUSTOMER_ID. A customer may be associated to several services, i.e. for each CUSTOMER_ID, there may be more than one SERVICE_ID. A customer is treated as having churned if there is a TERMINATION_DATE associated to each of the customer's SERVICE_ID.
- A win-back occurs when there is a COMEBACK_DATE associated to at least one of a customer's SERVICE_ID. The data set is not explicitly labeled. Churn and win-back information can be extracted from the data set in the following manner:

SUBMISSION

Competitors are required to submit the following through the competition website:

Problem 1:

Task	**Item to submit**
1	A table consisting of one column: The Customer ID.
2	A list of drivers that cause customers to churn. This list must be sorted with the most important drivers appearing first.
3	A table consisting of two columns: The Customer ID, and the main reason why that particular customer will not churn.
4	A table consisting of a single column: The Customer ID.

Problem 2:

Task	**Item to submit**
1.	Two tables, one for voice and broadband product, and another one for voice only. Each table: The Customer ID.
2	Two lists of drivers that cause customers to churn for the two products (voice and broadband, voice alone). Each list must be sorted with the most important drivers appearing first.

3	Two tables (one for voice and broadband, another one for voice only). Each table has of two columns, the Customer ID, and the main reason why that particular customer will not churn.

In addition to the above, competitors have to also submit a write-up (maximum 4 pages) describing the approach(es) that was taken to solve the two problems. This write-up must contain the names of the team members.

COMPETITION JUDGING

The sponsor of the data will use the churn and win-back data collected from 1 January to 31 March 2012 to judge the entries.
The precision measure will be used to score the submissions for Tasks 1,3 and 4 of Problem 1, and Tasks 1 and 3 of Problem 2.

Precision = tp / (tp + fp)
where tp stands for true positive and fp stands for false positive.

For Task 2 of Problem 1 and Problem 2, the data sponsor has internal list of drivers (the benchmark) created and ranked by its internal business intelligence system, and verified by the data sponsor's experts. The following explanation describes how submissions for Task 2 will be scored.

The score for a solution for Task 2 submitted by a contestant if computed as follows: where di is the driver at rank i as submitted by a contestant, and rank d(di) gives the rank of di in the benchmark list provided by the data sponsor. An entry that perfectly matches the benchmark list will have a score of 0, i.e. the best possible score for Task 2.

COMPETITION RULES

1) Anyone who complies with the rules of the PAKDD 2012 Data Mining Competition may participate. The organizers and the data sponsor are not allowed to participate.

2) Registration:
- All teams must register on the competition website in order to participate. To register a team, only the team leader needs to register.
- Each individual can only participate under one team. Multiple teams from the same organization are allowed with the condition that the team leaders are different.
- The names of team members must be disclosed when submitting the final entry. The composition of a team cannot be changed after the results are released.

3) Data are available for download only to registered participants.

4) Challenge duration: The challenge duration is about 4 weeks (16 April – 13 May 2012). To be eligible for prizes, competitors must submit their final entries by 13 May 2012. Each team is limited to one final entry. Only the last valid entry will be used for evaluation.

5) Reproducibility: Competitors do not have to deliver any code, but are required to produce a write-up (maximum 4 pages) to describe the methods that they have used to solve the problems.

6) Use of data: The data may only be used for this competition.

PRIZES

Prizes will be allocated as follows:
- Winner, Problem 1: MYR 1500

- Winner, Problem 2: MYR 1500
- Grand Prize Winner: MYR 2000

The winner for each individual problem is determined as follows:

- To be eligible for an individual problem's prize, a team must complete all tasks that constitute the problem. Contestants will be ranked for each task based on the scores computed for that task. The average rank is then computed for each team. For example, the average rank for a team for Problem 1 is (Rank for Task 1 + Rank for Task 2 + Rank for Task 3 + Rank for Task 4) / 4. The winner for Problem 1 will be the team with the team with the best average rank. The same principle applied when selecting the winner for Problem 2.

The Overall Winner is determined as follows:

- To be eligible for the grand prize, a team must submit complete ALL tasks in Problem 1 and Problem 2. The overall winner will be the team with the best average rank for all tasks.
- In the event where there is a tie, the judging panel with members from the organizer and data sponsor will decide on the tie breaker. Decisions made by this panel are final.
- Questions pertaining the competition should be directed to Lay Ki Soon at lksoon@mmu.edu.my.
- The prizes are sponsored by the Faculty of Computing and Informatics, Multimedia University.