Project Proposal

Team members: Boris Reznikov, Noa David, Ash Sirohi, Matt Hendren

File name: HR Employee Attrition – Updated Sample Dataset (Source: IBM Watson)

What is the exact business problem?

How to minimize employee attrition in the most cost effective way.

What might be the target variable?

The target variable will be the attrition rate, which is a binary variable (yes/no).

What is the use scenario?

The results of the study will inform what actions the HR can take to minimize employee turnover, and which types of employees to target. HR can then make informed, data driven decisions about how best to minimize employee turnover going forward. Another potential result is improving HR's hiring decisions; for example, we could predict that a certain type of person is more likely to stay with the firm if he/she is in a sales position vs R&D or if he/she needs to travel frequently or not.

Which is the data source?

The data source includes personal information about employees and job characteristics.

What precisely is the data mining problem?

One option would be to *predict* how long a certain type of person would stay with the firm given his/her personal characteristics. The other option is to do a causational analysis to understand if the company should change certain policies. For example, if the analysis shows that distance from home is a driving factor of attrition, the company could choose to only hire applicants who live within a certain distance of the office.

Data Mining Project Sketch
Section 101.Team - 103: Arturo Mora, Advyth Manepalli, Jayanth Barki

Business Problem & Use scenario:
We would like to use data from a sample of population from different areas of UK about their opinions of a certain artist, their level of interest in music and demographics (eg: age, working status, gender) to understand trends in music consumption.

This data can be very useful for music companies and apps like spotify. Whenever a music company is launching an album or spotify is suggesting an artist to a listener, they would need to answer the following questions for their marketing strategy.

1. Which segment of listeners (demographics, male, female) is likely to listen to the album/single?
2. How to reach the target segment with promotions for an upcoming album?
3. Are there geographic trends in Music tastes? If yes, then
    a. Can pubs and bars in an area play music in line with local tastes to attract more customers?
    b. Which artists would the local area buy tickets for a concert?
    c. Is it possible to targeting specific tours and promotional events within that area?
4. If a specific artist is looking to rebrand themselves, what is their current brand map.

Data Source:
Kaggle, EMI Music Data Science Hackathon:
https://www.kaggle.com/c/MusicHackathon/data

We are trying to get access to the 1 million interview data set, and if we are able to get access to it, then we would use it to mine through.
Data instance: Unit - We have in the dataset information about demographics, the average number of hours spent listening to music by the sample population, their rating as to how they discover music, willingness to pay etc. Unquantifiable variables include words that subjects have used to describe artists they were interviewed about.

Data Mining:
We will combine supervised and unsupervised methods to create the model to work on the data.

- Predictive analysis (how does an artist preference change with age, how does willingness to pay increase for age, music type, love for music etc.)
- Classification (market segmentation)
- Causal model (how likely is to like an artist given that you like another one)

Data Mining Project Proposal: Bike Sharing Program Predictions
Assignment DEC618 – Section 102 – Team 7
Justin Bedi, Leah Churinske, Alice Fan, Jianshi Yao

Summary:
By pulling bike rental data from kaggle.com (source: Capital Bikeshare), we are able to pull hourly rental data for a period of two years compared against weather data, and number of registered/unregistered users.

Note: the data only includes the first 19 days of each month over a two year period.

Business Problem

By analyzing the data, we can find the peak times of rental relative to temperature changes, holidays, time of year, etc. to better forecast demand, as well as when repairs and maintenance can occur to minimize service disruption.

Additionally, we can create a surge pricing model, charging slightly more based on time of day/year, and temperature.

Finally, depending on the results, we can determine whether this program can be rolled out to other cities across the globe, based on similar weather patterns.

What is the use scenario?

With the data, we can determine when the low and high demand periods occur, and when we need an influx of bikes available to users (both registered and nonregistered).

Which is the data source?

Kaggle (Completed Competition): https://www.kaggle.com/c/bike-sharing-demand

What is a data instance/unit?

Number of rentals, which help us predict peak/nonpeak times and whether this program can be operated in other cities

What might be the target variable?

Number of rentals on a given day/time based on temperature will be the target variable

What features would be useful? What precisely is the data mining problem? Is it supervised or unsupervised? How exactly would it add business value?

Being able to forecast bike rentals in different weather conditions using historical data, determining demand based on a specific day or time of day, or predicting whether this could be rolled out in different cities would be useful from a business perspective.

Currently, the data mining problem is that we may not have enough variables (currently only 12 variables available).