In BUDA530, we provided 2 project proposals and its methodology. BUDA535 course has given new insights to analyze data and reconsider and rethink the methodology of projects that we proposed. So far what we have learned in BUDA program, below are the revised methodology to investigate data on proposed projects by the team.

**Todd Erwin :**

For the sepsis model project, I think I would be able to introduce lasso to perform regularization and improve my variable selection. Currently I have been relying on subject matter experts to provide me with insight for predictors potentially matter. By using lasso, I could begin to explore various other predictors and possibly find variables that previously were not thought to have mattered and rule out some that according to opinion at this point do. This would empower me a little more than I had been previously to try some things then take the model to the SME and ask them if this made sense.

For the second project what I am really doing is trying to build a patient profile using population health data to predict readmission. I could leverage some of the clustering algorithms that we have discussed to place patients in different profile clusters than look to see which of those clusters tend to have a higher rate of readmission. What we could start doing is tailoring a patient’s care based on the profile to reduce the chance of readmission.

**Conor Frei :**

My first project was to see if I could predict a college basketball team winning a certain game based on a combination of team statistics. The goal would be to see which team statistics are driving certain team’s winning probabilities. Prior to BUDA 535, the method that would give insight on this information would be fitting a logistic regression model and then looking at the summary statistics. But after taking BUDA 535, there were a couple more ways to investigate the significance of variables that I would use for this project. The use of clustering and principal components would be two concepts that could elevate my dive into which team statistics are driving outcomes.

Since there are hundreds of team statistics that are calculated these days for college basketball teams, I can run a cluster function on all the predictor team statistics to see if any of them break teams out into defined groups that would give me more insight. Similarly, I can run principal components on the predictors to see if the data is telling me that some of the predictors can be left out in my calculations and if some are valued higher. I do not see a way to use the predicting functions into this type of project such as confusion matrix and glmnet because there isn’t a classifier type yes or no variable that can be applied to each team. Each game would be viewed differently so I would be looking for a percentage a team would win and not something that looks at past prediction history.

My second project is to look at if an NHL team pulls their goal in 3 on 3 overtime, would their odds of winning go up. Using data from 2015 to now, I was planning on using a logistic regression model to see if a team’s expected point totals would increase or decrease if they had pulled their goalie in overtime. I believe this one could also follow a little similarly to the additions to my previous project. I believe there would be some value in using clustering and principal components based on the variables used for each NHL team. I can attempt to use clustering to see how the data separates high winning overtime teams into groups and investigate their correlation of variables more in depth to see any underlying trends.

I believe principal components could also help me try to gain insight of trends in my data by seeing if any predicting variables could be dropped. If certain variables are showing as strong indicators for a team’s overtime performance, I can investigate a combination of those more with different weights to see if they better influence my model. As with the first one project, I do not see how the prediction functions covered in this class could be used for these projects as well. This type of project does not have a classifier variable that I could use that would give a binary option.

**Mona Patel :**

For my second project, Women (age 15-49) who are anemic (<11.0 g/dl), my goal is to analyze the significant factors associated in developing anemia and check the combinations of factors from these that are associated statewise. Previously, the method that i thought was to fit a binomial logistic regression model. But, after studying BUDA 535 course, I would use classification regularization method : elastic net and lasso model & classification trees and evaluate these methods that work best to investigate the significant factors and reduce the complexity if required as there are many variables in the dataset. Another part of the project is to see which combination of factors from these are involved statewise because factors for anemia varies region wise. For this part, I will be using clustering to investigate the relevant combinations of variables for developing anemia.

My First project of clickstream data analysis was already completed when I was working with IBM. For the next step which is to add more inventories based on bookings for peak seasons, customize offers, social marketing strategy, clustering can be a good method to check from groupings, bookings are done and based on these details, inventories can be added.

**James Smith :**

There are several red herrings in this data. The emphasis on change over time is irrelevant for an investment decision. As a result, I’ve decided to focus only on the Year 5 projections from GSMA. This eliminates the value of a time series analysis altogether.

The reason for focusing on Year 5, is that small changes in 4G density over time are less relevant than what the potential business landscape is later. Securing partners and then building a new network will take time. Therefore, it makes sense to understand the expected business environment at a given point in time.

There is the possibility that a new operator will spark rushed investment in the market and change the landscape. But given that most of the incumbents are state owned enterprises, we shouldn’t expect to see a significant jump in investments or upgrades to their networks.

Once the GSM Association (GSMA) data is combined with the population and tourism projections, then it would make sense to use a clustering method. The point being to identify multiple markets where an investment might be profitable.

Using clustering would support larger data sets allowing us to analyze potential markets beyond the Pacific Ocean to other underserved areas.

My second project is to Identify users abusing their cellular service package. This topic has become very important given the current stay at home orders issued by the Governor of Guam. Users have completely changed their usage patterns and overall data volume has increased 30%. We also suspect that revenue leakage has become revenue hemorrhaging as some specific users have increased usage as much as 90%.

While it’s possible to identify the current abusers, there are probably traits and warning signs that we could pick up to identify potential abusers before it happens. I need to modify the intent of this project to predict which factors might indicate a user is abusing their data package. Since these data sets can contain many variables, it might be an opportunity to use Bagging or Random Forests to reduce the model complexity.

**Greg Strohl :**

The essential question that I would like to answer is can we identify transactions that are potentially money laundering? For my project I would like to use data mining techniques to predict which transactions are potentially money laundering. Outlier detection is used in various applications such as fraud detection for credit cards and money laundering detection. I have learned several new methods in this class that I feel that could be beneficial in the detection of outliers. However, clustering appears to be one of the most obvious and effective means of detecting unusual transactions that would need to be further investigated.

I plan to use clustering as a means to create clusters of transactions that will be easier to isolate and analyze for money laundering. Clustering is a process that classifies the data into meaningful classes called clusters with the components of each cluster having the most similarities to each other and the members of each group have the least similarities to another group. The best performance of a clustering algorithm will be apparent when the clusters are distanced from each other as far as possible. In anti-money laundering, clustering is typically used for grouping transactions in different clusters that have the most similarities with each other. These techniques help us to detect patterns for suspicious transaction sequence or present models to identify the accounts or the riskier cardholders.

I feel that if I can determine if clustering is the best method in identifying money laundering, I can use this method to lower the number of false positives and make our investigation process more efficient.