**Mona Patel**

**BUDA 530 Group Discussion Project**

**Group Members’ Projects Synopsis :**

I first reviewed Greg’s career goals and project proposals. He has been working on investigating financial crimes for more than seven years. And with MS in Business Data Analytics, he wants to build upon his work and educational experience to find effective ways to fight financial crimes such as money laundering that he is already working on in his current role.

Based on his interests, his first analysis is to predict which transactions are fraud using Benford’s law using public data - Corporate payments of a West coast utility company 2010 dataset. He says that lower dollar amount transactions that slip by while detecting so I proposed if he can use normalization on amount variables that can help to create range. He is expecting to use a generalized linear model with splines or polynomials. His second proposed project was to predict the hurricanes and its significant factors using data science. But he changed and proposed the one that better fits his current job and can help benefit the organization. His second analysis is to predict the location where financial crimes will take place and significant factors for the financial crimes. He is going to use data from FINCEN database of SAR Filing Trend Data for years 2014 to 2018. He is expecting to fit a predicted model using non-linear transformations with variables available in the dataset and anticipate that similar models will predict at city, state and national levels.

James has been working in telecommunications for over thirty years and currently he is the Director of Wireless Networks at Cellular operator in Guam where he is responsible for network performance along with growth and development of the company. His goal is to become CTO of the company and direct the growth in all aspects.

Based on his experience, his first project proposal is related to his goal which is to identify the factors that would help improve investment opportunity for wireless and broadband services. He will be using data from several public sources including GSMA. While reading the background, I asked how he can come up with the number of customers, revenue per customer and cost of setting up telecommunication infrastructure. He will be fitting a predictive model to compare the current services and pricing available and predict to change the current market profitably and better serve the citizens of this Islands. His second proposed analysis project was to identify possible factors or combinations of factors that could explain higher instances of cancer in WV residents. But he changed it to the one that fits his experience in the telecommunication field. His second analysis will identify user patterns that might indicate either improper billing or user abuse. He will be using confidential customer data. He is planning to fit a generalized linear model to identify common traits for revenue leakage.

Conor is working as a systems consultant with Wells Fargo but his long term career plan is to move into the sports analytics field. With an increasing demand for data analytics across the major sports, he wants to use his previous experience and knowledge gained from BUDA program to work on data flows that helps make decisions to benefit the employers.

Based on his personal interest in sports, both the project proposals are related to the sports world. His first analysis is to find the relation between team categories which can predict the chances of a team to win the game. He is planning to fit logistic regression models with different categories as predictors and and win/no-win as response. His second analysis is to study if that is a smart strategy for the team pulling their goalie. He would be looking at the factors such as historical power play percentages and historical overtime records to see if this would create enough of an advantage for a team to try. Data will be collected from different sources and a logistic regression model would be used to look at the history data from year 2015 to forecast how many points a team would have gained or lost from adopting this strategy.

Stephen is promoted to manager shortly after enrollment into the BUDA program. He will be able to achieve his long term goal of reducing the cost of healthcare using analytics sooner by his additional leadership role which will allow him to make strategic decisions to benefit the organization. With the knowledge that he gained from BUDA program, he would like to create tools and models that can be implemented efficiently and effectively to reduce the cost of healthcare to both the providers and patients.

His first analysis is to increase the Case Mix Index to be within 80th percentile of the peer group. He is already working on this analysis for his current organization. After analysing inpatient claims data, observations were significantly below their peer group when billing for sepsis related DRG’s. A binomial logistic model is created using private data of current organization for the last 2 years of inpatient charges and performed cross validation to test this model. Review is in process and along with this, he is exploring other attributes such as physician notes to use to further refine this model. His second analysis is to identify a subset of populations who may be at risk of readmission based on social economic factors like access to transportation. A multinomial regression model with census data will be used to determine the influencing factors for a facility to have a higher readmission rate compared to peers with readmission rate as response variable.

**Group Feedback Synopsis :**

I am happy to receive feedback from all of my group members. Every member in our team is a great contributor to all the projects. Starting with feedback for my career goals, my team asked me to elaborate on my future goal and past experiences because I did not mention past experience and future expectancy in detail. While I was working in IBM, I worked on the first project Marriott Vacations WorldWide clickStream.

The main part of my first project for which I received feedback was to write more in detail on why POC was required and what business was looking for. And these comments were so useful to me and with regards to this, I wrote down the details of why and how we came up with the POC. Greg provided the best feedback with what factors are important in high conversion that helps the Marriott to expand that we found while doing this project. By determining the effectiveness of the site, Marriott could do a better job in marketing their resorts to travelers and can convert abandoned leads to opportunities. James emphasised on if the data is publicly available? After his comment, i mentioned in the project that this data is confidential and is clickstream data where sessions of the sites are used and customers details are used. Conor provided the feedback that I should add more details on what findings we did in the POC and what we came up with. I again added more details on our findings like the users’ pattern on the website, intervals between browsing and booking for vacations, checkout process for confusing and more details were asked in the checkout process. And with this, discussion with business, implemented the changes considering these factors and tested the POC and passed over the data to business for review. All these comments from my group helped me to expand a bit in every section.

First, I proposed a different project that I already worked on for my previous organization and it was to create a customer 360 portal for a Retail Bank. But I wanted to do data analysis related to healthcare with the knowledge that I gained from BUDA program till now so I changed the project #2 proposal which will benefit me with hands-on experience and utilize the skills that I am learning with this course.

My group had provided less feedback on my second proposed analysis project. Overall, they think that this project is good and will benefit me with my long-term goal to take into a health care employer interview to show my interest and research. They did think that predicting factors at state wise is great because anemia can be influenced by different factors in different regions and this will help to compare the influence factors for different states. James thinks that its reasonable expected outcome of this analysis.

**Career Goals :**

I have worked in different industries for over ten years with most of the time working in building web and desktop applications, data transformations , databases, APIs and dashboards.

I worked on .net technologies such as C#, web services(SOAP and rest API), hibernate and SQL server for over 6 years for different industries. While doing that, I worked with data to create reports using SSRS for clients that helps them to understand their business and make decisions. Later, there were times when data transformation was required to get data manipulated and fetch data from other sources. I learned MSBI during this time period to achieve the data transformation to dig deeper into the data. Used google analytics tools for few clients to help them understand the visitors and activities on the web applications.

For the last 4 years, I learnt Java technologies and its advanced framework, oracle database and ETL tool Pentaho. I worked as a full stack developer in IBM for Marriott Client where I spent most of the time working on data transformations and web applications. Along with this, created dashboards to provide information about vacation sites that are being booked during what period of time in the year. Based on this data, predictions and budgeting for the next 13 months were done. In addition, I worked on POC for clickstream analysis to analyze the clicks of vacation planning ads on social media and how frequent the ads are being clicked and successful booking done. While working with this client and understanding the application deeply, I realized that there are several opportunities in different ways to enhance the features by using data analytics tools that will benefit the business to grow. Some of the recommendations that can be easily listed are to increase the bookings of resorts, marketing campaigns based on most visited resorts, seasonal peaks, social marketing, sentiment analysis to know customer’s interest and their amenities requirements, add new resorts based on the most visited geographic locations.

These are the reasons why I decided to pursue an MS in Business Analytics to enhance my Data Analytics skill by learning Big Data technologies and machine learning. By leveraging these skills along with my previous experiences can help Organizations to make decisions to improve operational efficiency.

With the expanding growth of data analytics and machine learning across different industries, it will be important to be able to show my future employers how to leverage emerging data processing and machine learning techniques to make better decisions that would help them to serve their customers better, improve operational efficiencies. My short-term goal is to find a job as a data engineer in an organization but I am more interested in Healthcare industries which is a huge market for analysing data where I can think of a long term career.

Below are two projects, one of which is analyzing the clickstream data of one of the marriott web applications to understand consumers' interest for converting more leads into bookings and get the most visited resorts for advertisement campaigns. Another one is based on my interest in health care which is women who are anemic between age 15-49. I thought of this project because 3 of my close people had this problem and they were diagnosed with anemia when pregnant. Women at reproductive age , pregnant women and children are at high risk of developing anemia because of iron deficiency due to several factors.

**Project #1 Proposal** : **Marriott Vacations WorldWide (International) clickStream**

**Topic :** Identify the important factors to convert probabilities from the existing data to opportunities of the customers of Marriott Vacations WorldWide(International).

**Background** : Clickstream is the recording of areas of the screen that a user clicks while web browsing. As the user clicks anywhere in the web page, the action is logged. The log contains information such as time, URL, the user’s machine, type of browser, type of event (for example, browsing, checking out, logging in, logging out with purchase, removing from cart, logging out without purchase), product information (for example, ID, category, and price), total purchase in basket, number of items in basket, and session duration. This information can give valuable clues about what visitors are doing on your website, and about the visitors themselves. Clickstream analysis is useful for web activity analysis and market research. The navigation path can indicate purchase interests and price range. You can identify browsing patterns to determine the probability that the user will place an order. Data is of Marriott site and is confidential.

Marriott Vacations worldwide have multiple applications integrated together such as predicting and budgeting of vacation sites globally based on the existing trend in data, booking resorts, Inventory management and revenue system. Dashboards were created to visualize trends and predict for the future for the new sites to be created, quarterly revenues.

Consumers in recent years increasingly relied on the online medium for the travel research process. With this, Marriott Vacations wanted to add extra features by understanding the consumer’s interest and transactional patterns and use this data for social media marketing campaigns. I worked on POC taking these requirements into account with clickstream analysis.

**Goal** : Goals are to determine : 1. Marriott Vacations’s customers click on various resorts and effectively convert them into bookings if not done. 2. Geographic location of sites(resorts) having the highest count of clicks.

**Methodology** : Captured customer clicks, session details on a web server and streamed these details to Kafka Cluster. From the Kafka cluster loaded this information into ElasticSearch. Created a Kibana Dashboards to view the number of clicks on each resort and how many clicks converted into bookings and the geographic locations of top 10 resorts that got the highest clicks. Provided 50 observations to business for review to understand the reason for not getting converted to bookings. When discussed with business, reasons were mostly : The Checkout process is confusing, Site requires registration before purchase, No gift certificates , Changed mind and discarded cart contents, checkout requires too much personal information. Based on this, new requirements were provided by businesses to change the navigation of website and checkout process. These requirements are in production and are under observation.

**Outcome** : With this POC, business was able to understand the customer’s behaviour and what customers are looking for. The findings also reveal that a high percentage of abandonment occurs after searching for a hotel room and before submitting the final order confirmation. New requirements for changing navigation and checkout processes are under observation.

In addition, from most visited resorts, marketing was done through emails for now. This was still in process to go deeper to enhance the marketing based on clickstream data.

**Next Step :** Second goal was still in discussion along with more features such as personalization can be done on different customer touch points and add more inventories based on bookings for peak seasons, customized offers, social media marketing strategy to increase the conversion rates using the clickstream data and historical customer booking data were in discussion.

**Project #2 Proposal : Women (age 15-49) who are anemic (<11.0 g/dl)**

**Research/Topic :** Significant factors associated with the risk of developing anemia in women of reproductive age or when pregnant.

**Background :** I propose to analyze which significant factors are associated in predicting anemia (mild, moderate or severe) among women of reproductive age and pregnant women state wise as some of the states are having a high anemic rate in women. I will be using public data provided by NFHS (Open Government Data(OGD) Platform India). The NFHS programme is conducted by the Ministry of Health and Family Welfare and implemented by a group of survey organizations and Population Research Centres(PRCs) to provide estimates of important indicators on family welfare, maternal and child health, nutrition and other health issues. The NFHS used the following six questionnaires to collect data: household questionnaire, woman’s questionnaire, man’s questionnaire, biomarker questionnaire, fieldworker questionnaire, and the verbal autopsy questionnaire.

My interest in this project arose especially because few of my friends and family were diagnosed as anemic between second and third trimesters. They had to take extra supplements of iron and vitamins along with prenatal vitamins during this period of time and always worried about premature delivery. Because maternal anemia is associated with maternal and child morbidity and mortality such, as increased risk of miscarriage, stillbirth, prematurity, and low birth weight of the baby. There are many significant factors that need to be considered such as age, socio-economic issue, residential elevation for predicting anemia. On the other hand, this project might help me professionally to find a job in the healthcare industry at some point.

**Goal** :The public data from NFHS is nationwide and I plan to analyze the significant factors involved in developing anemia among women of reproductive age and pregnant women and predict the anemic or no-anemic in them at state level.

**Methodology** : NFHS has data from 2016 to 2017 and is about to publish data for 2018 and 2019. The data consists of 699,686 observations nationwide. I will download the data in CSV and import it in GoFirst and will need some cleanup. Data is nationwide so I plan to analyze these at the state level and compare the predictors for states with west states (my hometown and border states) to see if same variables are involved in west states and other states. I will use a binomial logistic regression model with response variable as anemic or not. Predictors variables possibly be age, education level, household wealth status, sources of drinking water, state, area of residence - urban or rural , fertility status, age at first birth, total number of children ever born, adverse pregnancy outcome for last/most recent pregnancy, birth in the past five years, breastfeeding status at the time of the survey, and iron folic acid for 100 days or more when they were pregnant, Postnatal and prenatal care.

**Expected Outcome** : The outcome of this study will be the presence of anemia among women and also the factors associated with the risk of developing anemia among women of reproductive age and pregnant women between age 15-49. I anticipate that this model will predict results and validate if the same or different factors are associated statewise.

Link : <https://data.gov.in/>

**Draft Version :**

Below is the draft that I sent it over to my group for feedback. I changed my project #2 because that was already done and i want to do the analysis of health data related to health care.

**Career Goals :**

I have worked in different industries for over ten years with most of the time working in building web and desktop applications, data transformations , databases and dashboards.

For the last 4 years, I spend most of the time with data transformations and web applications for Marriott Vacations worldwide project in IBM. Along with this, created dashboards to provide information about vacation sites that are being booked during what period of time in the year. Based on this data, predictions and budgeting for the next 13 months were done. In addition, I worked on POC for clickstream analysis to analyze the clicks of vacation planning ads on social media and how frequent the ads are being clicked and successful booking done.

When I decided to pursue a MS in Business Analytics at West Virginia University, I wanted to enhance my Data Analytics skill by learning Big Data technologies and machine learning. While I was working with IBM on clickstream POC, I realized how data analysis can help Organizations to make decisions to improve operational efficiency.

Below are two projects, one of which is analyzing the clickstream data of one of the marriott web applications to understand consumers' interest for converting more leads into bookings and get the most visited resorts for advertisement campaigns.

**Project #1 : Marriott Vacations WorldWide (International) clickStream**

**Background** : Clickstream is the recording of areas of the screen that a user clicks while web browsing. As the user clicks anywhere in the web page, the action is logged. The log contains information such as time, URL, the user’s machine, type of browser, type of event (for example, browsing, checking out, logging in, logging out with purchase, removing from cart, logging out without purchase), product information (for example, ID, category, and price), total purchase in basket, number of items in basket, and session duration. This information can give valuable clues about what visitors are doing on your website, and about the visitors themselves. Clickstream analysis is useful for web activity analysis and market research. The navigation path can indicate purchase interests and price range. You can identify browsing patterns to determine the probability that the user will place an order.

**Goal** : Goals are to determine : 1. Marriott Vacations’s customers click on various resorts and effectively convert them into bookings if not done. 2. Geographic location of sites(resorts) having the highest count of clicks

**Methodology** : Capture customer clicks, session details on the web server and stream these details to Kafka Cluster. From the Kafka cluster load this information into ElasticSearch. Create a Kibana Dashboard to view the number of clicks on each resort and how many clicks converted into bookings and the geographic locations of top 10 resorts that got the highest clicks.

**Outcome** :

**Project #2 : Customer 360 Platform for Retail Bank**

Background : Every Retail Bank needs a smart platform where they can have a 360 degree view of their customers. Customer 360 platform provides valuable insights into Customers Spending behavior, connected payments, product details and connectedness with other customers.

**Goal** : Create Customer 360 portal for a Retail Bank which can provide insights into customers spending behavior, product holding, connected payments done by customers (top 10 debit and credit parties), how customers are connected with other customers. This portal can be leveraged to users by Customer Relationship managers, product managers to understand their customers, offer different products accordingly.

**Methodology** : Used Big data technologies like Spark, Sqoop to extract customers data (credit/debit cards transactions, online transactions, product holding, customer profile etc) from various sources and transform it. Load transformed data into NoSQL and GraphDB(e.g. Neo4j, AWS Neptune). Create a Dashboard or Portal which would fetch customers data from NoSQL and GraphDB.

**Outcome** :

**Project #2 Proposal : Women (age 15-49) who are anemic (<11.0 g/dl)**

**Research/Topic** : Significant factors associated with the risk of developing anemia in women of reproductive age or when pregnant.

**Background** : I propose to analyze which significant factors are associated in predicting anemia (mild, moderate or severe) among women of reproductive age and pregnant women state wise as some of the states are having a high anemic rate in women. I will be using public data provided by NFHS (Open Government Data(OGD) Platform India). The NFHS programme is conducted by the Ministry of Health and Family Welfare and implemented by a group of survey organizations and Population Research Centres(PRCs) to provide estimates of important indicators on family welfare, maternal and child health, nutrition and other health issues. The NFHS used the following six questionnaires to collect data: household questionnaire, woman’s questionnaire, man’s questionnaire, biomarker questionnaire, fieldworker questionnaire, and the verbal autopsy questionnaire.

My interest in this project arose especially because few of my friends and family were diagnosed as anemic between second and third trimesters. They had to take extra supplements of iron and vitamins along with prenatal vitamins during this period of time and always worried about premature delivery. Because maternal anemia is associated with maternal and child morbidity and mortality such, as increased risk of miscarriage, stillbirth, prematurity, and low birth weight of the baby. There are many significant factors that need to be considered such as age, socio-economic issue, residential elevation for predicting anemia. On the other hand, this project might help me professionally to find a job in the healthcare industry at some point.

**Goal** :The public data from NFHS is nationwide and I plan to analyze the significant factors involved in developing anemia among women of reproductive age and pregnant women and predict the anemic or no-anemic in them at state level.

**Methodology** : NFHS has data from 2016 to 2017 and is about to publish data for 2018 and 2019. The data consists of 699,686 observations nationwide. I will download the data in CSV and import it in GoFirst and will need some cleanup. Data is nationwide so I plan to analyze these at the state level and compare the predictors for states with west states (my hometown and border states) to see if same variables are involved in west states and other states. I will use a binomial logistic regression model with response variable as anemic or not. Predictors variables possibly be age, education level, household wealth status, sources of drinking water, state, area of residence - urban or rural , fertility status, age at first birth, total number of children ever born, adverse pregnancy outcome for last/most recent pregnancy, birth in the past five years, breastfeeding status at the time of the survey, and iron folic acid for 100 days or more when they were pregnant, Postnatal and prenatal care.

**Expected Outcome** : The outcome of this study will be the presence of anemia among women and also the factors associated with the risk of developing anemia among women of reproductive age and pregnant women between age 15-49. I anticipate that this model will predict results and validate if the same or different factors are associated statewise.

Link : <https://data.gov.in/>