**Applied Data Science Portfolio**

**Syracuse University**

**By Samuel Rogers**

[**Sroger07@syr.edu**](mailto:Sroger07@syr.edu)

**Table of Contents:**

The following is a crosswalk which identifies the program learning objectives achieved while also identifying the projects and associated page numbers utilized to showcase these learning objectives. While I have numerous impressive projects I would be proud to present, these specific projects incorporate multiple learning objectives. I found it more impressive to showcase the combined learning objectives in a few projects, than explain numerous projects with singular learning objective. Links to the GitHub files can be found by clicking the project titles below.

1. Describe a broad overview of the **major practice areas** of data science.
   * Determined via the collective use of these projects presented
2. **Collect and organize** data.
   * [*IP Prediction* – IST 707](https://github.com/samuel-rogers-data/IP-Prediction)
     + **Pages 3-6**
3. **Identify patterns** in data via visualization, statistical analysis, and data mining.
   * [*IP Prediction* – IST 707](https://github.com/samuel-rogers-data/IP-Prediction)
     + **Pages 3-6**
   * [*How to Live Forever* – IST 687](https://github.com/samuel-rogers-data/How-to-Live-Forever/tree/main)
     + Pages 6-10
4. **Develop alternative strategies** based on the data.
   * [*How to Live Forever* – IST 687](https://github.com/samuel-rogers-data/How-to-Live-Forever/tree/main)
     + Pages 6-10
5. Develop a plan of action to **implement the business decisions** derived from the analyses.
   * [*Process Improvement: Check Reconciliation* – MBC 638](https://github.com/samuel-rogers-data/Process-Improvement-Check-Reconciliation/tree/main)
     + *Pages 10*-12
6. **Demonstrate communication skills** regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
   * [*Process Improvement: Check Reconciliation* – MBC 638](https://github.com/samuel-rogers-data/Process-Improvement-Check-Reconciliation/tree/main)
     + *Pages* 10-12
7. Synthesize the **ethical dimensions** of data science practice (e.g., privacy, Unfair Discrimination, Reinforcement of Human Biases, Lack of Transparency).
   * [*How to Live Forever* – IST 687](https://github.com/samuel-rogers-data/How-to-Live-Forever/tree/main)
     + Pages 6-10

***IP Prediction* – IST 707**

**This project identified the initial production volume of an oil and gas well, using numerous parameters of the well, prior to producing oil from the reservoir in production phase. Traditionally, a reservoir engineer uses neighboring wells up to 2-3 miles away to predict the initial production volume. This isn’t always very accurate, as the ground makeup can change drastically from one area to the next.**

**The largest and most challenging task for the IP Prediction project was collecting and organizing the three SQL databases. These sources came from public and private sources such as DrillingInfo, FracFocus and Oseberg dataStream. The Oseberg database was connected to my company’s Azure database which also served as the originating dataframe. In this case it was the left table for a left join to the other two databases. FracFocus originated from a public government site which had numerous additional columns not necessary for the project. The raw file I had to begin with contained 4.5 million observations. The data added was primarily the Proppant and Fracture Treatment attributes (Table 1). DrillingInfo is a paid site which provided many of the missing or incomplete attributes which Oseberg lacked. The database was filtered to include only horizontal, actively producing oil wells from Grady County Oklahoma in SQL. This was necessary for the preliminary analysis as after joining the databases in R Studio based on API, there were a combined 283 columns. This was reduced to 10 highly deterministic attributes seen below (Table 1).**

**Table 1**: Highly deterministic data terms

|  |  |
| --- | --- |
| **Attribute** | **Description** |
| API | Unique identifier of well |
| Operator | Company serving as overall manager of drilling the well |
| Reservoir | Subsurface body of rock containing hydrocarbons able to be produced |
| Prod\_Type | Type of hydrocarbons being produced |
| MD | Measured depth is the distance traveled underground (ft) |
| Completed\_Interval | Portion of horizontal well open and producing from the reservoir (ft) |
| Fracture\_Treatment | Chemicals, fluid volume, proppant quantify used in fracture (unstructured format) |
| Pumping\_Flowing | Production method of oil and gas from reservoir |
| Lateral\_Length | Length of horizontal well (ft) |
| TVD | True Vertical Depth is distance from surface to deepest point (ft) |
| Frac\_Fluid\_W | Fluid injected under high pressure into hydrocarbon bearing rock to increase flow of oil and gas (gallon) |
| F\_m\_BOE | First month oil and gas quantities converted to equivalent energy produced by barrel of crude (barrel) |
| Proppant | Sand used to keep rock fractures ajar (lbs) |

**Oil & gas data is notoriously messy. Rows with a single NA tended to have numerous additional NA values. Any NA values led to the entire row being excluded, except regarding oil or gas volumes. A well can produce a single product or both products. These NA values were set to zero and a composite value was created based on standard petroleum conversion practices called F\_m\_BOE, which stands for first month barrels oil equivalent. This is the y-output. Next the numeric F\_m\_BOE was distributed into equal frequency bins and set as categorical variables (Table 2). The y-output is converted and uses an 80-20 random split between train and test data. This is done so it can be predicted upon using Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes and Gradient Boosted Machine (GBM).**

**Table 2**: Discretized First Month BOE separated into bins of equal frequency

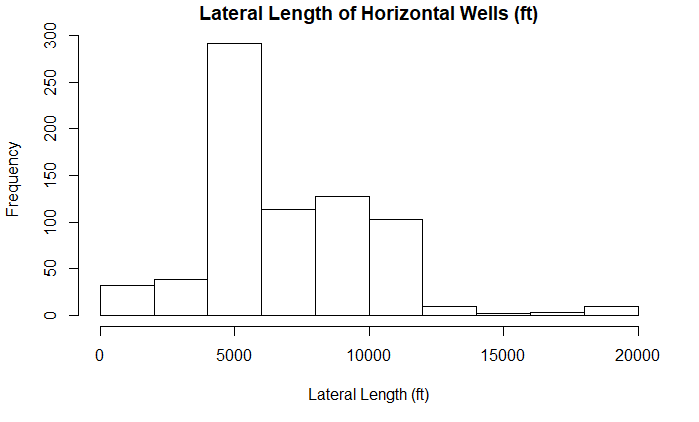
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **First Month BOE (barrels)** | | |  |
| **Trickle** | **Low** | **Medium** | **High** | **Super Max** |
| **2 – 1,350** | **1,350 – 4,670** | **4,670 – 10,000** | **10,000 – 18,100** | **18,100 – 69,400** |

**Diving into the lateral lengths of the wells, some interesting insights surrounding the clusters of wells drilled were found. Typically, when drilling a horizonal well, 1, 2, or 3 miles are purchased for development. This means the expected lateral length should cluster around 5,280, 10,560 or 15,840 feet. Instead, there was much less clustering around the mile markers than expected, as can be seen in Figure 1. While still present at 1-mile, it was anticipated that 2-mile and 3-mile marks would have greater clustering. The lack of clustering can be better identified in Figure 2. Additionally, there was no relationship between longer lateral wells and greater initial production volume. Logically more surface area from a longer well would produce more oil, however this wasn’t the case for initial production.**

Chart, scatter chart

Description automatically generated

**Figure 1**: Production volumes vs. lateral length do not have a positive linear relationship as might be expected when there’s more surface area to extract petroleum from.



**Figure 2**: Lateral length have high frequency at 1 mile (5280 ft), but no discernable clustering at the other mile markers (10,560 and 15,840 ft) .

**The best results on accuracy were achieved from the Gradient Boosted Machine using 20-fold cross validation and a 200 tree, 5 node size, Random Forest model with 38.5% and 37.6% respectively. While the accuracy was poor at predicting which of the 5 bins were expected to produce oil, the specificity was above 80% for the bins and the predictions tended to predict close to the correct bin (Figure 3). In bringing quantifiable accuracy to a value which before simply averaged neighboring wells, is an improvement. Figure 3 supports that the GBM is useful in predicting close to the correct bin range, if not the correct one.**

Text

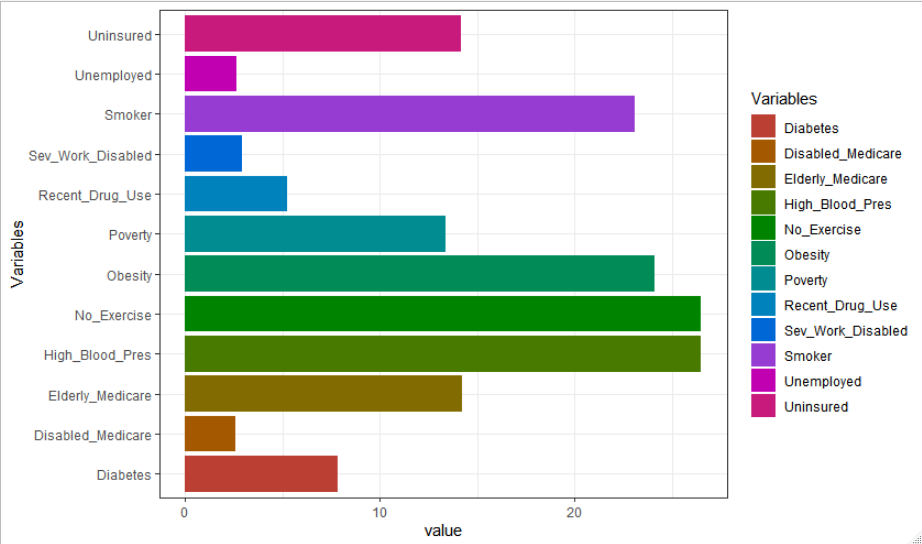
Description automatically generated

**Figure 3**: Gradient Boosted Machine reference vs prediction are close to correct bins.

**This project developed my skills in numerous aspects of the program’s core objectives. While acquiring and cleaning data for the IP Prediction project, my ability to join databases, convert data types, decipher optimum data sources, and determine appropriate use cases for supplementing missing data was strengthened. Identifying patterns in the data via visualizations, statistical analysis and data mining were strengthened as well. This was best shown while discovering the oil well lengths do not fall into crisp 1-mile increment clusters. Lastly in using multiple data sources to collect data across an area, machine learning techniques were utilized to provide new insights about the data, not previously seen in the data.**

***How to Live Forever* – IST 687**

This was a group project which examined the relationship between United States health statistics per county and average life expectancy (ALE). Data was collected from the Community Health Status Indicators to combat obesity, heart disease and cancer (CHSI). The analysis was conducted in R and R studio primarily. After acquiring the dataset, the initial task consisted of narrowing the 573 unique columns from each US county into a manageable 140 variable database. This was completed using a series of read\_csv and left joins for the 7 files, excluding numerous erroneous columns in the process. The top risk factors effecting ALE are shown in **Figure 4.**



**Figure 4**: Leading risk factors contributing to premature death

To create the **Figure 4** bar chart, I used *melt* from reshape library to sum all the counties and express them as a total value per variable. Lastly, I pivoted the axis, and applied an electric rainbow color palette to the chart. This was an important figure as it succinctly showed the top candidates causing a low ALE across the country. This was important in seeing that *Poverty* was not a major indicator of ALE, however in later studies a strong correlation between *Smoking*, *High Blood Pressure* and *Lack of Exercise*, thus we considered *Poverty* a co-variable and thus still an important part of our analysis.

The ALE across the nation was categorized and can be seen in **Figure 5**. This was broken up by state but there is also an R Shiny app which provides the ALE and top offending variables. This utilized two scripts, one for the user interface controlling the layout & appearance while the other is a server script to house the directions for recipient computers to build the app themselves. If a specific county has high values in the realm of *Lack of Exercise*, the user could learn that they should work out more than their fellow neighbors as their county is statistically lacking and thus leading to their collective lower life expectancy.

Map

Description automatically generated

**Figure 5**: Average Life Expectancy per U.S. county

Another alternative analysis of the data can be shown in **Figure 6.** This graphic tells a more specific story about the effects smoking and lack of working out have on ALE. Using GGplot and by reversing the x-axis and setting limits for the smoking range, it can be seen clearly how the data supports that more work outs and less smoking lead to greater ALE.

Chart, scatter chart

Description automatically generated

**Figure 6**: Correlation between Smoking, not working out and ALE

In this next visualization, poverty across the United States was the focus. **Figure 7** shows poverty in the US on a state level. The US map is imported using map\_data(“state”), and poverty per county is melted to the state level. Next by matching the database’s state names to abbreviations, the original database can be matched to the US map. For clear distinction, the poverty levels were cut into 6 intervals such that they can represent the percentage of the population which is impoverished. **Figure 7** show that neighboring states have similar poverty levels, meaning poverty is not evenly distributed but disproportionate across the United States.

A picture containing map

Description automatically generated

**Figure 7**: Neighboring counties and states tend to have similar levels of poverty percentage

While there are numerous census variables the team could have used, race was something we decided early on to exclude from our correlative analysis. This was based on an ethical responsibility we felt we had as data scientists to avoid reinforcing human biases in a sensitive subject area. Ultimately this helped lead us to our broader theme of identifying high risk behaviors which could be changed to lead a longer, healthier life.

With such a large dataset and many routes for interpretation, we investigated numerous alternative theories when studying this subject area. As mentioned earlier, the topic of race was excluded, as it could have gone against our ethical duties as data scientists. Because of this we developed alternative strategies to focus on behaviors which were ethical and could theoretically be changed[[1]](#footnote-2). Another example of an alternative approach taken was poverty. While it didn’t directly cause a low ALE, it had a high correlation with many high-risk behaviors which lead to premature death. Initially poverty was a major part of our investigation until we developed alternative strategies based off the data. Data visualization uncovered patterns in the data suggesting neighboring states shared similar poverty levels. This might be difficult to detect with just a table, however when creating a visualization, it becomes more apparent. Statistical patterns are visualized in the data by showing a strong negative correlation with ALE to both smoking and not working out. All in all, by data mining census data about people all across the United States, our group was able to identify larger over-arching patterns and co-correlations like poverty, which shorten live expectancy.

***Process Improvement: Check Reconciliation* - MBC 638**

This next project analyzed methods to increase the efficiency and reduce systematic errors in processing checks paid to Cosmo Energy. By defining the problem, implementing measuring systems, utilizing multiple analysis techniques, and implementing improvement, the process was controlled and able to sustain the process reduction by 25% and lower annual costs by $7,200.

*Problem Statement: Company wastes spending ability by inefficient deposit log system. Engineers hindered by slow deposit system and missing data to meet monthly revenue deadline.*

After defining the problem statement, a measurement system (**Table 3**) for the project was established to quantify errors being made, whether that be in excessive time, duplicated data, or missing data.

**Table 3**: Measurement system to identify errors in the system to improve upon.

Table

Description automatically generated

The system was then measured over the course of 3 months using excel version control software. Using the rules established in **Table 3**, the measured errors accumulated by each client over the 3-month observation period are displayed using a pareto chart **(Figure 8).** This was created using built in excel graphics.

The benefit of **Figure 8** is that it shows the total errors created and which ones are causing the most trouble. In this case, it was determined that a handful of clients were causing a significant portion of the lag time and errors. They were consistently causing issues and it was due to the way in which the client notifies Cosmo Energy that they have paid and there is information to acquire. As part of my improvement phase, I set up direct deposit notifications from the bank to a shared email address of the employees involved in the process.



**Figure 8**: Pareto Chart identifies company checks which create the most errors and are left unprocessed the longest

The final process can be seen in **Figure 9**. This improved process reduced the total steps required, consolidated duties to individuals instead of a shared responsibility and incorporates direct deposit instead of relying on physical bank runs.

Diagram

Description automatically generated

**Figure 9**: Improved process resulting in a 25% reduction in processing time.

This project taught me how to systematically develop a plan of action in a business setting and implement those actions based on an analysis. The plan was developed utilizing the DEFINE, MEASURE, ANALYZE, IMPROVE, CONTROL presenting cycle. This presentation of analysis was via the pareto char, then the process was improved upon in the process map. In completing this project for the betterment of Cosmo Energy, I was also able to improve my own abilities in communicating my analysis to a broad range of business stakeholders, rather than just fellow data scientists. By showing a clear plan with a 0.2 in Sigma Quality level, the people involved, timeline, and methods for continued success, my managers were able to visualize the benefit and grant me additional processing control. This method of explanation was also useful in communicating with the other engineers, accountants, and data input employees involved in the project. Upon completion of the project I presented my findings to our CEO, who was happy with the work completed and results achieved.

**Conclusion**

The following course projects were utilized in demonstrating the required learning objects required for completing the Applied Data Science portfolio at Syracuse University: IST 707, IST 686, and MBC 638. Presenting these works in a concise essay has in its own way showcased my abilities to communicate data analysis to a vast audience of business professionals. Nearly every project at Syracuse University has required my collection and organization of data, however *IP Prediction*—IST 707 was the most challenging due to the immense size and disarray of the databases. Data visualization was particularly useful to showcase the lack of a grouping pattern in *IP Prediction*—IST 707 well lengths. The project itself utilized a massive amount of data which I was able to extract insights and statistics from which were not previously visible.

The visualization of poverty levels in the United States in *How to Live Forever*—IST 687 helped to show neighboring states have similar poverty levels. This project also showed using statistical analysis the leading causes of a low life expectancy. By data mining census data, the group was able to uncover numerous correlations and patterns in this massive database which when combined together, showed co-correlations such as poverty, which relate to other major factors in decreasing average life expectancy. These weren’t previously visible in the data prior to this deeper analysis. How to Live Forever also required alternative strategies to the data. Initially poverty was suspected as a major effector of ALE. Later in the project our analysis suggested it was only co-correlated, and highly correlated to variables leading to low ALE. Initially race was a focal point of our analysis and while our data source provided in depth detail with correlations we discovered, we felt it violated our data science ethics to avoid promoting data which reinforced negative human biases surrounding. This helped us to narrow our project to focus on behaviors that could theoretically be changed such as exercising and not smoking.

The final project discussed *Process Improvement: Check Reconciliation* – MBC 638 which utilizes a systematic method for presenting a business case to stakeholders which will addresses a process which can be improved upon. In later steps the process is measured and analyzed until finally improvements are implemented to create the success promised in the business case. By working on this project my company was able to clearly see the intended output for my boss and expectations of the employees involved in my process improvement.

Completing this Applied Data Science Masters at Syracuse University has been one of my proudest accomplishments. I finally feel that I have found an industry as diverse as myself and one which allows me infinite possibilities to apply my knowledge to systematically improve processes and find new data insights through machine learning.

1. We realize that some of these attributes such as poverty, disability or diabetes are not a life choice and potentially impossible to change, while a smoking, lack of exercise or eating more vegetables are more realistic life choices one could incorporate in their lives. We did not mean to offend anyone in the making of this study. [↑](#footnote-ref-2)