Portfolio Project: Option 2

Jessica Ham

MIS530- Predictive Analytics

Colorado State University – Global Campus

Vincent Agboto

January 5, 2020

Portfolio Project: Option 2

1. **BUSINESS UNDERSTANDING**
   1. **Business Objective-** The challenge, as it is stated, is that a hospital is seeing the same viral symptoms in patients each fall but does not have a process by which to predict the extent of this issue or to react by preparing with enough medication to support the patients. The hospital has asked for help developing a model that can determine the volume of medicine to keep on hand at hospital and the costs associated with over or under stocking during the fall season.
   2. **Assess Situation**- Presumably the hospital has extensive records on patients that have been treated for this virus in the past, such as whether they came in multiple years with the same issue (recurring), whether it passed through people working in the same place or living in the same household (how it is transmitted), symptoms and conditions surrounding the people getting ill (environment). By collecting data on patients both with and without the virus from the past several years, we should be able to feed this into a model to understand the population which is contracting the virus. Further review will need to be done to understand, of those that have the virus, who is seeking treatment with us, with other medical facilities, or possibly who is not seeking treatment at all. This information is important to consider when we look at medication that needs to be kept on hand for our facility.
   3. **Determine Data Mining Goals**
      1. Understand the characteristics of people in the area contracting the virus- The first step is the building of a model to understand who is getting the virus and what characteristics they have that are different that those in our area who remain healthy. Is the environment a factor? How does the virus spread?
      2. Convert those characteristics into an estimate of people contracting the virus in the area- With a working model based on the population of past patients in our hospital, we can apply this model to the area population at large to get a read on the number of people who are contracting this virus in our metro area. This is important to consider as there are people who in the past may not have come to our hospital for treatment and could this year. We need to be sure we know the full possible extent of the population.
      3. Understand the likelihood of those people in the area coming to our hospital for treatment- Finally, understanding the population who has come to our hospital for treatment in the past, we can create a secondary model that attempts to estimate of the infected population, what percent are likely to come to our hospital for treatment and what is the cost of over or under supplying medication for that population? We need to make sure that we find the most profitable balance to treat the most patients and have the least shortage or overage of medication.
2. **DATA UNDERSTANDING**
   1. **Collect/Describe/Explore Data**-
      1. Review hospital records for last 3 years to create a training data set. It is important to have patients that were and were not diagnosed with the virus for model training. It may be necessary to over sample the virus infected population in order to get a strong insight into the factors that attribute to the diagnosis. The variables for our model should include symptoms, but also meta data such as address, work/school location, age, etc that might help determine transmission or environmental factors.
      2. Collect meta data for the area. Our secondary model is going to try and understand what portion of the infected population is coming to our hospital and how much medication do we need to keep on hand for them? The patient meta data collected above may also be useful in this model for understanding what patients do/do not come in for treatment. Other questions that could be asked: What is the population area for patients that come to your hospital? How many other facilities are in the area that may have had patients? Why did those patients choose another facility and would it benefit us to try and attract those patients to our hospital?
   2. **Verify Data Quality**-
      1. In creating the secondary model, it will be vital to understand how much of the area population comes to this hospital as opposed to others or who might refuse treatment. We need to make sure of the quality of information that we collected about the area to get this part right. An ideal model for this would have information on other hospitals in the area, what they are charging and whether they had enough medication on hand in prior years as well as information on the people that went to those hospitals. Obviously the details of those patients will be unavailable for our model, so this is where the quality of our population meta data will be key. Average incomes for various areas and ages etc can possibly be imputed from details we have on our hospital patients and public records to create a good population for this secondary model.
      2. Another data quality concern will be to ensure that the population we consider positive for the virus is clearly covered in the data collection. If we estimate that only about 5-10% of our total patients within our hospital were diagnosed with this virus, this may not be enough data points to successfully model with. We want to make sure we have a clear idea when collecting and ensuring quality of data how this infected population compares to the uninfected. We may choose when selecting data to oversample to ensure a good model, but first we must understand the population as it is.
3. **DATA PREPARATION**
   1. **Select Data**- Once we have collected all of our data, the next step is to select patients in a way that ensures the population we consider positive for the virus is represented well in our training data set. As mentioned above, the population is too small to get a good read. If only 10% of our patients contracted this virus, then in a training set of 50%, we may only get 5% or less who have the virus to train from. “If you do not balance the number of instances, most classification algorithms will heavily focus on the majority class” making it seem you are getting results, but actually possibly underperforming on the minority class, which should be the focus (Lukas, 2019, para 7). To correct this, we should start by taking all patients that were positive for the virus into our data set, and then fill in the remainder of our sample with non-viral patients at a ratio closer to 50:50, to ensure a good representation of both populations. “This intentional sampling process, designed to incorporate more (typically low-prevelance) members of a certain community into your sample is called oversampling” (Vaughan, 2017, para 4).
   2. **Clean Data-** Once you have oversampled, it is important to weight your data to maintain the percentage of total patients accurate to the real ratio. This side effect of oversampling is known as a selection bias, which occurs “…when the criteria used to recruit and enroll patients into separate study cohorts are inherently different” (Pannucci, 2011, para 8). In our case, selecting for people within our population with the virus and balancing it against an equal group without the virus. But, now that we have our population, it is important to remove the implicit bias that our oversampling has created. “The main correction technique used…consists of reweighting the cost of training point errors to more closely reflect that of the test distribution” (Cortes et al, n.d., para 3). We are using these extra patients to understand the characteristics of the virus and those with it, but we do not want to overstate their prevalence in the population. With this in mind, one tool that could be useful for such a model creation, evaluation and deployment is SAS Enterprise Miner. Right in the data selection step, one can reweight the population to reflect the original ratio of viral to non-viral patients, removing this bias from the results while still keeping the oversampling to improve a viral prediction.
   3. **Construct/Integrate/Format Data**- Without having the data available to actually conduct this experiment, this section is difficult to envision. Once we have selected the data with care, we can review the formats for consistency, review missing data for decisions on inclusion, rejection or imputing. Considering the manner in which the data was selected, it would be ill advised to reject any virus positive candidates, so in those cases, missing data would need to be imputed. The options here are an impute node to infer all missing information for a regression model or to create selection criteria within the Decision Tree model node that allows data to be presumed or ignored as it passes through the tree.
4. **MODELING**
   1. **Select Technique**- In reviewing articles for this assignment, I found many that discussed models that are most successful in hospital evaluation environments. Newz Kesornsit did not post his research specifically, but did blog that he found when looking at multiclass and binary classification techniques for readmission of diabetes patients, reducing the data to a binary classification and then performing a decision tree technique provided the best performance (Kesornsit, 2019, para 1). Several journal publications I found most useful included Zolbanin and Delen’s article on improving outcomes for readmissions through processing of electronic medical records (Zolbanin, 2018) and Araz, Olson and Ramirez-Nafarrate’s article on hospital admissions to the ER using triage information (Araz, 2019). Both seemed particularly relevant to this hospital’s predictive analysis modeling needs and both found results that decision trees outperformed regression models for classification of this nature. While Zolbanin found his best result in a gradient boosted tree, Araz found a random forest was his best performer, commenting on how “…as the number of observations used in the training and validation of the model increases XGBoost model’s performance increases while others stay constant or perform worse” implying that the model has more scalability that the other tested models as well (Araz, 2019, section 6).

Additional research published in *PLoS One* cited a decision tree as an effective method for predicting the severity of Dengue Fever patients in Malaysia and suggested using it as a tool to “…differentiate mild from life threatening infection…” to avoid costly and unnecessary hospitalizations (Tamibmaniam et al, 2016, para 1).

* 1. **Generate Design**- With this in mind and the prior discussion of the data available and questions to be answered, I propose a two pronged model which begins by running the patient population through a decision tree algorithm to group individuals who are more likely to contract this virus and their characteristics. With a trained decision tree, you could run the population at large which the hospital treats through the tree model to get a feel for those who are most likely to contract the virus. Finally, to answer the question of quantities and cost of medication as well as profit, I propose a linear regression technique to understand how that relates to our hospital’s patient count and therefore how much medicine we will need.

It is understood from the business objective that the real goal here is to get an answer for how much medicine to order for the upcoming fall based on the history. The issue is that we don’t have a clear idea on who is contracting the illness and as shown from the prior studies listed above, a linear model is not a good performing model for that type of classification. Before we can determine how many orders of medication are needed, we need to understand the profile of the virus positive patient. To some, this would answer the question of what to order- number of cases treated at the hospital equals number of doses needed. I would propose however that there are several places to get treated in the area and that if one was looking at maximizing profit for the area, the question would be whether we could easily attract clients from other locations. Further, what about people who don’t seek treatment? Why don’t they? Could we offer a discount or something else that might attract more clients? Once we know what that patient looks like, we can predict with some accuracy the number of cases in our general geographic area, rather than just our prior patients, by simply running the full area population through our trained and validated model. This count of viral cases could be fed into a traditional linear regression model, along with sales price information to the patient and information on the cost of treatment to the hospital. This could create a profit model to understand the costs of over ordering medicine and trying to attract new clients versus ordering just based on the prior client counts and running out if new clients arrive naturally.

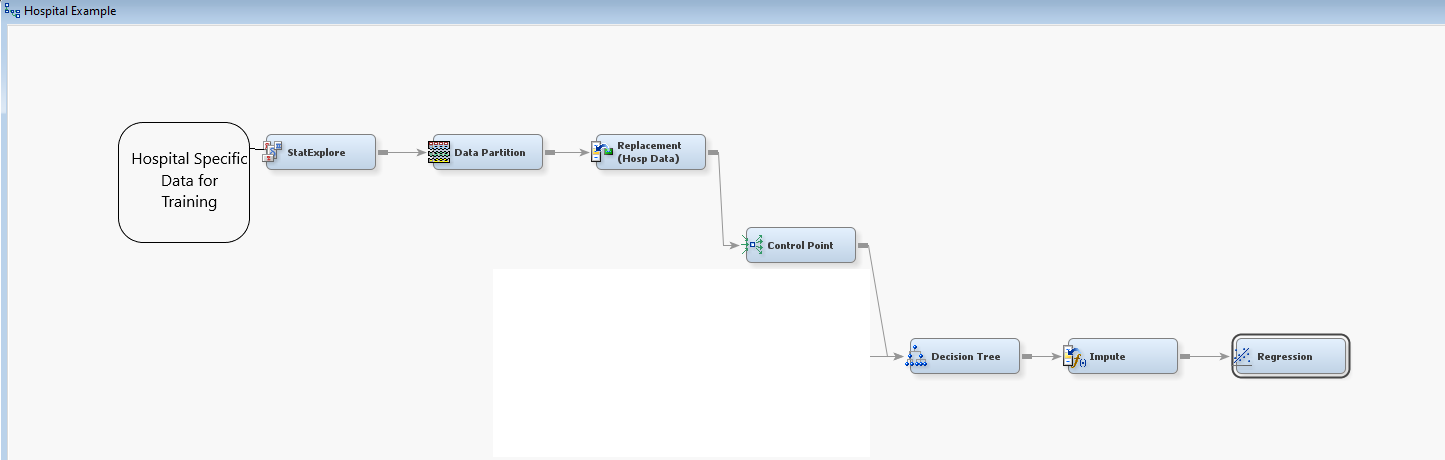
* 1. **Build & Assess Model –** The process to actually build and assess this proposed model would best be suited to SAS Enterprise Miner as it is one of the easiest tools for feeding the results of one model into another through the diagram tool. Without having actual data to load into the system, it can be challenging to visualize what that would look like, but I would imagine a diagram chart similar to Figure 1 shown here: 

Figure Possible Diagram Layout

As indicated by the model, the results from that decision tree node would feed directly into an impute node to fill in missing values for the final regression model. It would also be possible to impute values on the front end so that the same data fed through both models, rather than letting the decision tree make one value assumption and then later create a potentially different decision moving into regression. It may be however that the values used in each model are different and the values you are imputing were not needed prior to the regression model.

1. **EVALUATION** 
   1. **Evaluate Results-** With our Model running in two stages, the evaluation will be slightly more complex than a single model. The first priority will be ensuring the accuracy of the decision tree model, since it is going to feed the regression model down the road. Some metrics that can be useful when reviewing the performance of a classification problem are: accuracy, precision, recall, F1 Score, and Receiver Operating Characteristic (ROC) curve.
      1. **Accuracy** in a sentence is the “…number of correct predictions on the total number of predictions” (Kakkad, 2018, para 10). In our case, that is out of all the predictions from the selected training data, what percent were accurately predicted as sick from our virus or as not sick from our virus.
      2. **Precision** is only the positive responses. The percentage of our patients that we predicted correctly as sick versus the total patients we predicted as sick (whether they were or not).
      3. **Recall** is similar but focuses on the percentage of times we predicted our patients as sick and they were actually sick.
      4. **“F1 Score** is needed when you want to seek a balance between precision and recall”, with the understanding that a balance between correct sick predictions may be best (Shung, 2018, para 13). It could be that our disease is highly contagious and missing a positive means 10 new positive cases? Or maybe the medicine is very costly to the hospital, and the false positive means a loss of profit on a disease that isn’t that contagious? This score attempts to take into account that there is no one best measure of a good model as it depends on the situation. As we are attempting to maximize profit, this measure is useful, especially as we move forward asking a question about clients getting medication elsewhere.
      5. **ROC curve** plots the False Positive Rate (times we said the patient was sick but they were not) on the X-axis, with the True Positive Rate (times we accurately predicted a sick patient) on the Y-axis. This is plotted against a linear plot of performance if patients were randomly declared sick or healthy of our virus. The Area Under the Curve (AUC) is a measurement of performance compared to random. The more area under the curve, the better the performance of your model compared to random selection.
   2. **Review/Determine Steps-** If our Decision Tree has a good ROC curve and meaningful F1 score to our needs, we would be able to move forward into the regression model created from that data to determine the most profitable point for ordering the medication. The accuracy of a regression model can be determined in a few ways, but the most common is an R-squared value, “…which is the proportion of variation in the outcome that is explained by the predictor variables” (Kassambara, 2018, para 2). If we are happy with the model’s ability to predict profit, based on sick patients and costs associated with attracting them, we would consider deployment.
2. **DEPLOYMENT**
   1. **Plan Deployment** – Planning Deployment of a successful model would involve planning how we will update the source data to include the full area population, as opposed to just the population we trained and validated with, which included only patients of our hospital and was oversampled and then weighted to more accurately focus on the viral positive patients. This modification should not impact the performance of the model, but does involve making sure the data points that are used in the model have been scrubbed to match the data set which was originally used.

Further the question of who has access to this model and when and how will they use it must be answered. As the data scientist for the hospital, are you simply running the model on your own and reporting results back to the administrators for dissemination? How often will this report need to be updated? These questions are factors to be considered before deployment.

* 1. **Plan Maintenance** – The other side of deployment to consider is the maintenance of the model. The understanding of the Business and Data, preparation of Data, Modeling, Evaluation and Deployment are steps known as the CRISP-DM (Cross Industry Process for Data Mining). This method is designed as a structured approach for designing and maintaining any data mining or analytics process. This iterative process allows that during maintenance, there could be a need to re-evaluate any step in the process, as data inherently will change and the model must be able to adapt to those changes. Considerations for this phase might be how often should the accuracy of the model be reviewed and by whom or how often the data itself is to be updated and what are the processes needed to manage the input?
  2. **Produce Final Report & Review (Conclusion)-** As mentioned in the business understanding, though it is vital to understand the viral diagnosis, the real output here would not be the number of patients predicted, but rather an estimate on how many of those patients our hospital believes it should strive to treat while maximizing profit. This is another area where the SAS Enterprise Miner system is useful. When loading data and adding the weights to correct our selection bias, there is also an area to assign value to the yes and no answers, creating a profitability matrix that the model can use to assess the best answer to our specific question. The profit of treating a sick patient can be entered, as well as the profit or cost of falsely treating a healthy patient or failing to treat a sick patient.

With this matrix in place, Enterprise Miner has the ability to plot our models based on the decision matrix and the profit figures at each patient decision. This would be the final report, as it will support a decision about at what point we should stop treating patients for the sake of profit or how many patients that are not ours that we wish to attract. As much as we would like to be ready to treat everyone, the reality is that a hospital must be profitable and ordering medication that is wasted or turning away patients because not enough medication was bought are both means to reduce profit. The balance must be struck and this is the model that will support that.

References

Araz, Ozgur M, Olson, David and Ramirez-Nafarrate, Adrian (2019). Predictive Analytics for

Hospital Admissions from the Emergency Department Using Triage Information, *International Journal of Production Economics, February 2019;* 208:199-207. Retrieved from: <https://www-sciencedirect-com.csuglobal.idm.oclc.org/science/article/pii/S0925527318304778>.

Cortes, Corinna, Mohri, Mehryar, Riley, Michael and Rostamizadeh, Afshin (n.d.). Sample

Selection Bias Correction Theory. Google Research and Courant Institute of Mathematical Sciences, New York, NY. Retrieved from: <https://cs.nyu.edu/~mohri/pub/bias.pdf>.

Frei, Lukas (2019). *A Deep Dive Into Imbalanced Data: Over-Sampling.* Retrieved from:

<https://towardsdatascience.com/a-deep-dive-into-imbalanced-data-over-sampling-f1167ed74b5>.

Kakkad, Ruchir (2018). *Model Selection by Performance Comparison.* Retrieved from:

<https://medium.com/@RuchirKakkad/model-selection-by-performance-comparison-8f5912f8bd86>.

Kassambara, Alboukadel (2018). Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp

and more. Retrieved from: <http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/>.

Kesornsit, Newz (2019). *Reducing Multiclass to Binary for Classification Techniques of Re-*

*Hospitalization of Diabetes Patients.* Retrieved from: <https://medium.com/@newzkesornsit/reducing-multiclass-to-binary-for-classification-techniques-of-re-hospitalization-of-diabetes-cb7f975a22cb>

Pannucci, Christopher and Wilkins, Edwin G (2010). Identifying and Avoiding Bias in

Research, *Plastic and Reconstructive Surgery 2010 Aug;* 126(2):619-625. Retrieved from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917255/>.

Shung, Koo Ping (2018). Accuracy, Precision, Recall or F1? Retrieved from:

<https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>.

Tamibmaniam, Jayashamani; Narwani Hussin; Wee Kooi Cheah; Ng, Kee Sing; Muninathan,

Prema (2016). Proposal of a Clinical Decision Tree Algorithm Using Factors Associated with Severe Dengue Infection, *PLoS One;* San Fransisco, 11(8). Retrieved from: <https://search-proquest-com.csuglobal.idm.oclc.org/docview/1813621858/fulltextPDF/2677A455F56F4A4FPQ/1?accountid=38569>.

Vaughan, Roger (2017). Oversampling in Health Surveys: Why, When, and How?, *American*

*Journal of Public Health, 2017 August;* 107(8): 1214-1215. Retrieved from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5508166/>.

Zolbanin, Hamed M. and Delen, Dursun (2018). Processing Electronic Medical Records to

Improve Predictive Analytics Outcomes for Hospital Readmissions, *Decision Support Systems, August 2018;* 112:98-110. Retrieved from: <https://www-sciencedirect-com.csuglobal.idm.oclc.org/science/article/pii/S0167923618301076?via%3Dihub>