The University of Texas at Austin

MIS 382N 11: Business Intelligence Capstone

Improving Population Health:
Risk Stratification within CommUnityCare
Final Report

---

by

Arjun Adapalli, Siqi Chen, Corey Haines, Korawat Tanwisuth

Sponsors

CommUnityCare, UT-Dell Medical School Department of Population Health

Instructors

Dr.Michael Hasler, Dr.Ramesh Rajagopalan

April 9, 2018

Contents

Executive Summary

Hospital readmission is one of the costliest expenses facing hospitals in the United States. Readmissions can be an implication of improper or inadequate treatment for inpatients and result in expensive penalties from the Centers for Medicare and Medicaid Services. By correctly identifying high risk patients for readmission, pre-discharge interventions and effective follow-up care management could be performed, ultimately leading to significant reduction in readmission rates. This project focused on a developing a risk stratification model for CommUnityCare based on five-year patient records from 2013 through 2017. Predictive models were built using demographic information, historical admission records, and geography-based socio-economic data to predict whether or not the patient will be readmitted in the future. Based on the performance of our predictive model, it was estimated that $386,100 could be saved per one thousand patients in CommUnityCare services. Lastly, model predictions were explained by important risk factors contributing to each individual encounter. Knowledge of readmission risk factors prior to discharge can help help improve the coordination of post-discharge care between care management teams, thereby leading to higher quality treatment and higher profits for CommUnityCare.

1 Business Background

1.1 Business Context

One of the largest cost drivers in U.S. healthcare on both the macroeconomic scale as well as organizational level is costs associated with readmissions. These costs fall on both the public and private sector, with Medicare paying for 58 percent of all readmissions in 2011, private insurance paying for 20 percent, and Medicaid covering the remaining 18 percent [1]. In 2011 Medicare spent $15 billion alone (and 37 percent of its total expenditures)[1] on inpatient admission and readmission costs. To combat these high costs and reform healthcare in the United States, the Obama Administration passed The Affordable Care Act (ACA) in 2010. Section 3025 of the ACA added section 1886(q) to the Social Security Act, establishing the Hospital Readmission Reduction Program (HRRP)[2] to combat readmission problems by requiring CMS to reduce refund amounts awarded to IPPS hospitals with excess readmission. CMS defined readmission as an admission to an acute care hospital within 30 days of discharge from the same or different acute care hospital.

Since going into action on October 1, 2012, Medicare costs for readmission have risen to $27 billion dollars, with $17 billion spent on readmissions that could be classified as potentially avoidable [3]. Furthermore, penalties associated with these costs have risen to $528 million as of 2017 [4]. Due to increasing costs Medicare and Medicaid look to enact even more stringent policies towards hospital readmissions. CMS measures hospital performance under the HRRP by calculating excess readmission ratios (ERRs) for each of the six program measures: acute myocardial infarction (AMI), heart failure (HF), pneumonia, chronic obstructive pulmonary disease (COPD), coronary artery bypass graft (CABG) surgeries, and elective primary total hip and/or total knee arthroplasty (THA/TKA) . An ERR is the ratio of predicted-to-expected readmissions for a given measure. Under the current non-stratified methodology, measures with 25 or more eligible discharges and an ERR greater than 1.0 enter the payment adjustment factor formula. An ERR greater than 1.0 indicates that a hospital performed worse than the average performance of all hospitals. The payment adjustment factor formula is used to calculate the size of the payment reduction. For fiscal

year (FY) 2013 through FY 2018 the payment adjustment factor was calculated as follows [2]:

$$1 - min\{0.03, \sum_{dx} \frac{Payment(dx) \times max\{(EER(dx) - 1.0), 0\}}{Allpayments}\},$$

where dx is one of the six program measures.

However, beginning in FY 2019, hospital performance will be judged relative to the performance of hospitals within the same peer group. Hospitals will be stratified into five peer groups based on the proportion of dual-eligible stays, i.e., the proportion of Medicare fee-for-service (FFS) and Medicare Advantage stays where the patient was dually eligible for Medicare and full-benefit Medicaid. The median ERR of hospitals within the peer group will be used as the threshold to assess hospital performance on each measure rather than an ERR of one. Doing this will minimize the potential to disproportionately penalize hospitals serving indigent and impoverished populations. Furthermore, a neutrality modifier (NM) is applied to scale payment adjustments in order to retain a similar amount of Medicare savings under the stratified and non-stratified methodologies. For FY 2019 and beyond the payment adjustment factor will be calculated as follows [2]:

$$1 - min\{0.03, \sum_{dx} \frac{NM \times Payment(dx) \times max\{(EER(dx) - MedianpeergroupEER(dx)), 0\}}{Allpayments}\}$$

Because hospital performance will now be judged relative to their peers, it is imperative for them to take aggressive measures to drive down their readmission rates. Achieving this would decrease the payment adjustment factor and consequently amount to substantial savings. However, one of the problems hospitals face is not knowing which patients are at high risk for readmission prior to being discharged from the hospital. Thus, current intervention methods such as discharge re-engineering and transition care management, which rely heavily on whether hospitals can identify these high-risk patients before prior to discharge, have only made marginal improvements in overall readmission rates.

1.2 Business Problem and Opportunities

CommUnityCare, Inc. is one such organization concerned about the rising costs of hospital readmissions. Providing services at 19 locations in Travis County, CommUnityCare serves

approximately 88,000 individual patients in the greater Austin area per year. This includes inpatient and ER services as well as outpatient primary healthcare, dental care, limited specialty care, lab testing, radiology, a full service pharmacy, and behavioral health services. These services are provided to all Travis County residents including those whose incomes and lack of private health insurance qualify them for enrollment. CommUnityCare strives to achieve health equity for all by: (1) being the health care home of choice; (2) being a teaching center of excellence; and, (3) providing the right care, at the right time, at the right place. However, meeting this standard can be challenging given the patient population they serve.

Many of CommUnityCare's patients are from low income areas with higher poverty rates and lower education levels. This results in a higher utilization rate for CommUnityCare's ER's rather than outpatient and primary care services. Rather than driving to outpatient facilities, some patients will use the ER as a one-stop shop for all of their healthcare needs. However, from CommUnityCare's perspective, this is a misuse of their resources, and if a patient returns to the ER within 30 days, CommUnityCare's readmission rate increases. Furthermore, due to financial constraints, impoverished patients tend to have lower compliance rates [5], thus making them higher at risk for readmission. Therefore, CommUnityCare faces a greater risk of having preventable patient readmissions than other healthcare providers within Travis County.

For this reason, our project with CommUnityCare and Dell Medical School focuses on building a risk stratification model for CommUnityCare's patient population. The goal is to identify high-risk patients so that actions may be taken prior to discharge that can help circumvent preventable readmissions, thereby driving down CommUnityCare's payment adjustment factor which will result in reduced costs and increased profits. Successfully identifying patients with high likelihood for readmission could translate to improved quality of care and significant cost savings given that an inpatient readmission costs roughly $12,300 for those with Medicaid, $13,800 for those with Medicare, and $14,200 for those with private insurance [1]. Additional savings would be accrued through higher utilization rates of the organization's constrained resources.

1.3 Conceptual Framework and Objectives

More specifically, our capstone team set out to develop a risk stratification model tailored to CommUnityCare's patient population that assigns each patient a risk score for readmission, classifies those patients with high readmission risk, and identifies significant factors that contribute to their risk score.

It is important to first stratify CommUnityCare's patient population because not all segments will have the same risk profile. Moreover, these segments may utilize CommUnityCare's resources in different manners. Thus, it is critical to identify these moieties to provide CommUnityCare with a better understanding of the diversity within their patient population, and also help them tailor their treatment strategies to different segments.

From here, a classification model can be built that will predict whether a patient will be readmitted after a visit to one of CommUnityCare's facilities. Based upon the predicted probability of being readmitted within 30 days, each patient will receive a risk score upon discharge. Additionally, it will be important to know what factors contributed to a risk score at the patient level. This level of insight will hopefully enable CommUnityCare to take any precautions prior to discharge and take necessary follow-up measures to prevent the readmission.

The last objective for our final deliverable is to synthesize all of the above information into a final dashboard product that can be integrated into their current infrastructure. We envision that this dashboard will present the risk score as well as risk-contributing factors for each patient to nurses, doctors, and other healthcare providers at CommUnityCare. This project will conclude with a final presentation to the Research Board at CommUnityCare describing our workflow and the business value of the project to their organization.

2 Technical Methodology

2.1 Data Description

The dataset provided by CommUnityCare covered a five-year period covering the dates 01/01/2013 to 10/19/2017. The composite primary key, i.e., unique identifier, in the raw

dataset is person ID, hospital visit date, and diagnosis code. Hence, each row represents a patient encounter with a single international classification of diseases (ICD) diagnosis code. However, a single patient could have multiple diagnosis codes for a single encounter as well as multiple diagnosis sequences. A diagnosis code describes the condition that is chiefly responsible for the admission of the patient to the hospital, whereas the diagnosis sequence includes not only this information but also information regarding comorbidities and/or secondary diagnoses. These rows were eventually rolled up to the encounter level (see pre-processing) so that each row represented a single patient encounter with all relevant diagnoses.

Also included in the dataset were features relating to patient demographics. These included ethnicity, uniform data system (UDS) ethnicity, UDS race, UDS homeless category, address, zip code, city, county, age, and language. Since Austin is such a highly stratified socioeconomic city, we supplemented the dataset with other social determinants of health (see feature engineering) that could influence risk for readmission.

Additionally, the data contained information related to the care site, care management, and length of stay. These features include the hospital place of contact, encounter type (inpatient vs. ER), healthcare provider, payor name, last CommUnityCare location, last CommUnityCare visit date, and discharge date.

Lastly, information about the patient's LACE score was supplied. The LACE index is a well known index which identifies a patient's risk for readmission or death within thirty days of discharge. One could view the LACE as a simple "predictive model". It consists of a checklist filled out by the healthcare professional seeing the patient and covering four areas: Length of stay, Acuity of admission, Charlson comorbidity index, and Emergency department visits within the last 6 months. The scale ranges from 1-19, with 0-4 representing low risk, 5-9 representing moderate risk, and greater than 10 representing high risk for readmission. Considering the fact that this metric is low-cost and relatively trivial to calculate for each patient, we aim to have our model outperform LACE in terms of predictive ability. Otherwise, CommUnityCare could just default to the LACE index for assessing risk for patient readmission and avoid computer costs associated with running machine learning algorithms.

2.2 Pre-Processing

In order to have a dataframe suitable for running a classification model, it was necessary to first clean the raw data. The first step in this process was to drop duplicate records, as these redundant entries would lead to inaccurate counts for the number of historical patient visits as well as overcount the number of readmissions in the dataset. Another redundancy issue with CommUnityCare's data was the presence of records with the same patient ID, same hospital visit date, and same diagnosis but different LACE scores. This could have been due to differences between subjective evaluations of a patient encounter or errors in data entry. Because it is impossible to determine which LACE score is the most accurate, we used the last record for a patient encounter and dropped all other records with the same patient ID, hospital date, and diagnosis code.

After handling duplicate records, we rolled up the data to the encounter level. To do this we used patient ID and hospital date as the composite primary key to identify a unique row. We collected all diagnosis codes and diagnosis sequences associated with an encounter and stored them in a list. The diagnosis information was then removed from the encounter table, and used to create a new diagnosis table. Having a separate table for diagnostic information helped us later on with deciding which diagnoses to include as dummy variables in our classification models.

Starting October 1, 2015, CMS issued an update to their ICD coding system and transitioned from ICD9 to ICD10 for Medicare and Medicaid compliance. Because the dataset includes admissions before and after this event, both ICD9 and ICD10 codes were present in the dataset. We used a crosswalk to convert ICD9 codes to their appropriate ICD10 counterparts. Not doing this step would have led to additional noise in the dataset and a potential reduction in accuracy of our classification model. Although the diagnosis codes were now consistent across the time period in the dataset, there were still tens of thousands of unique diagnosis codes and sequences. Because only 384,238 patient encounters existed, we faced a dimensionality concern. To reduce the number of new columns that would need to be created to represent each unique diagnosis code, we mapped ICD10 codes to their clinical classification software (CCS) 10 codes. This step reduced the number of dummy variables

for diagnosis codes from several thousand to about 150.

One last issue with patient encounters was missing data for the discharge date column. For missing values we assumed the patient was dismissed from the hospital on the same day as admittance, as per the input from the Dell Medical School sponsors. This was a strong assumption to make and later influenced our values for new variables such as length of stay and time between consecutive visits.

2.3 Feature Engineering

After cleaning the dataset, we proceeded to create new columns derived from the existing columns in the dataset. These include counts of past ER and inpatient visits within the last 12 months, the previous visit type, interval between the current visit and previous visit, next visit type, interval between current visit and next visit, as well as total number of visits within the last 12 months. Subsequently, we created indicator columns for transfers-in and transfers-out. A transfer-in or transfer-out represent transfers between facilities on the same day, i.e., the interval between the previous visit and current visit is less than one day or the interval between the visit and next visit is less than one day. These instances needed to be flagged because they would have been counted as readmissions otherwise. Instead, they represent hospital triage and redirection of patient care to the appropriate facility.

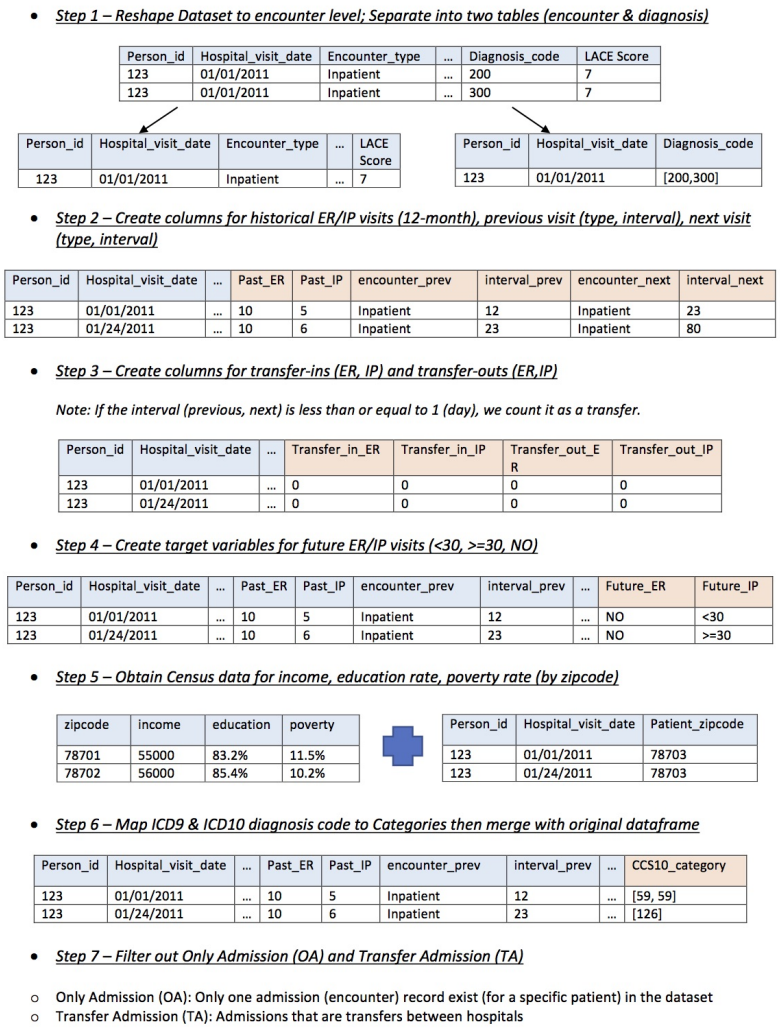
From the interval between the current visit and next visit we constructed our target variable. We determined whether a readmission occurred within the next 30 days (as per the CMS definition), and we also considered whether they came back to any of CommUnityCare's facilities within the next year. The logic behind considering readmissions beyond the 30 day CMS window is that these admittances still levy substantial costs to hospitals through utilization of resources. Although these admittances would not count against their payment adjustment factor score, they would still contribute to higher costs and lower profit margins per patient. Therefore, we looked at all readmissions within a 3 month period.

After constructing our target variable, we scraped data on education levels, poverty rates, and median income for all zipcodes found within the dataset. Previous work has shown that social determinants of health such as education level can have an effect on hospital outcomes and readmission rates[6], so we decided to try and capture some of these determinants for

our analysis.

The last steps taken in our pre-processing and feature engineering phase were to drop one-time admittances and transfer admissions from the dataset based on 3M methodology. A patient with only one admission during the timespan of the dataset would likely introduce noise to our models. As stated previously, same-day transfers should not count as readmissions since they are part of the triage process. Pictured below is a summary figure of our pre-processing and feature engineering workflow (figure 1).

Figure 1: Workflow for pre-processing data and subsequent feature engineering



2.4 Model Selection

For our modeling process, we focused on the high-cost readmissions for CommUnityCare. To do this, we broke our dataset into three different subsets depending upon the next encounter of a patient: inpatient to inpatient, ER to inpatient, and ER to ER. Since overnight stays cost significantly more for hospitals than same-day ER visits, we decided to spend a majority of our time training models for inpatient to inpatient and ER to inpatient instances. However, if a patient repeatedly shows up to the ER beyond what is expected in a years time frame, then they could be misusing CommUnityCare's ERs as their one-stop shop primary care center. Ideally, CommUnityCare would want to reroute these patients to their outpatient clinics, which are designed to handle more routine checkups at a much lower cost than an ER visit.

For each subcategory we started our classification modeling process with simple logistic regression models. We did not expect these linear classifiers to have the most predictive power, but they do provide more interpretable results than more complex models. Therefore, we thought these models may provide insight into which predictors to feed into the more complex tree-based methods and neural networks. These more complex methods generally do well in terms of predictive ability because they can find interactions not readily seen in the data. Because they assume less conditions than generalized linear models, they also tend to be more robust when encountering new data points.

To compare model performance, we used the area under the curve (AUC) metric from the receiver operating characteristic (ROC) curve. The ROC curve is a graphical illustration plotting the true positive rate of a binary classifier against the false positive rate for a binary classifier. We used AUC as our performance metric because it takes into account both false positives as well as false negatives. Metrics such as accuracy only look at the proportion of correct predictions and would not be a reliable metric to evaluate models in our project, since there is a large class imbalance in our target variable.

3 Findings

3.1 Exploratory Data Analysis

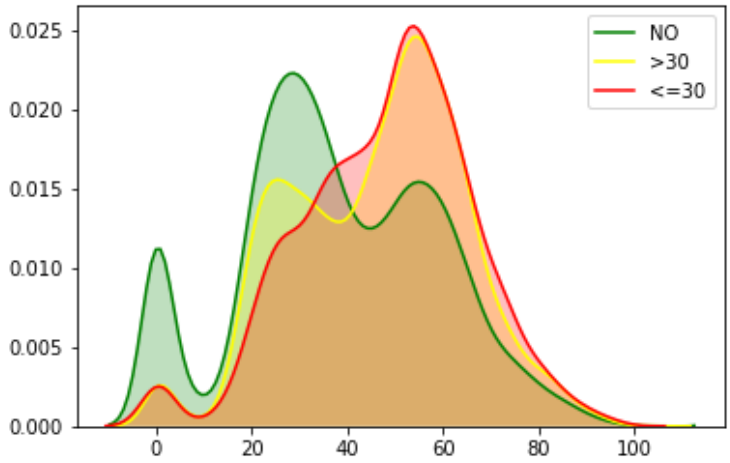
Our analysis focuses mainly on inpatient readmission, a readmission that includes patients who are admitted as inpatients and get readmitted again as inpatients within a certain time window. While this type of admissions constitutes roughly 10 percent of our dataset, the business value gained from preventing this type of readmission is significant.

In our exploratory data analysis we mainly focused on discovering the underlying information and patterns hidden in the data. We computed several demographic statistics to help us better understand our target population. Interestingly, we found that the majority of our patients are middle-aged Hispanic and Latino individuals located in Travis County (around 45%). Using zip code information, we incorporated socioeconomic factors such as the average income, education level, and poverty level from census data. The CommUnityCare patient population has an average income of $54,000/yr and a median income of $48,000/yr, which means the income distribution is right-skewed. The education column gives a percentage of the population that has completed high school, while the poverty column tells the percentage of the population that is considered to be in poverty. While the demographic information provides more insight into our population, it still does not tell us its complete risk profiles. Looking at the distribution of past inpatient visits, we see that the 50 quantile is at 95 days, meaning that 50 percent of our patients are readmitted within 95 days. To further understand the risk profile, we partitioned patients into three categories: readmission within 30 days, greater than 30 days, and no readmission. This choice of partitioning is based on how readmission is defined by the Center for Medicare and Medicaid Services (CMS). Also, hospitals will be penalized mainly for readmissions that are within 30 days. Thus, this will help us focus on the right group of patients from a business perspective.

We examined the target variable class distributions of each predictor attribute to determine if any of these attributes provides a clear separation across the target variable classes. We hypothesize that such attributes will contribute to higher prediction accuracy. We see that the age distributions for the classes differ in shape (figure 2), which suggests age might

be a good feature to include in our model. The plot also suggests that older people are more likely to be readmitted.

Figure 2: Distribution of age based on readmission status



Similarly, we can infer that if a person is not homeless, then the probability of readmission is lower (figure 3).

Figure 3: Distribution of homeless status based on readmission status

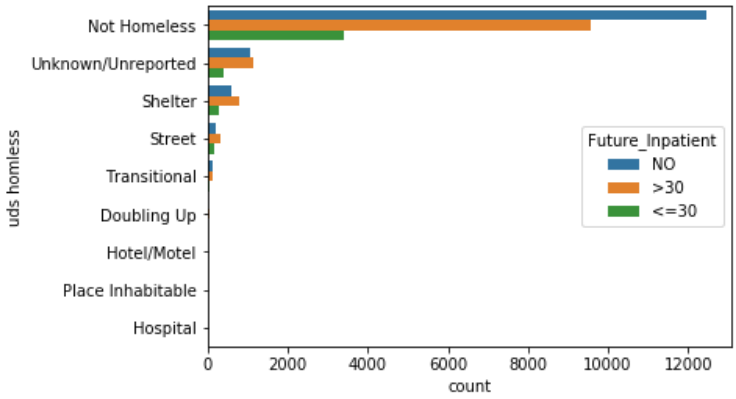
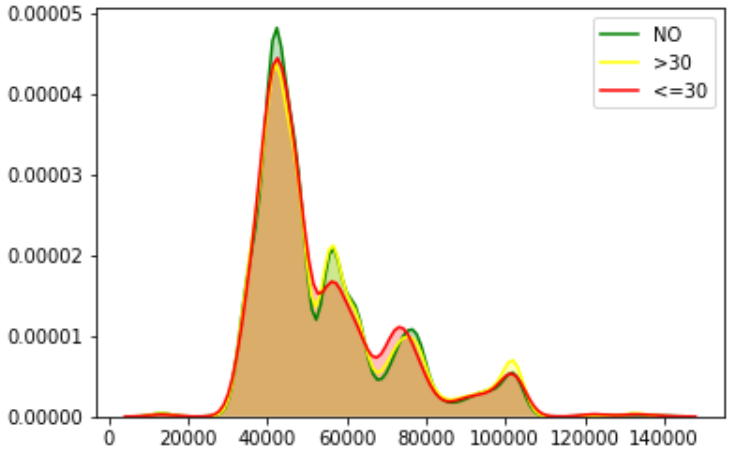


Figure 4 shows an example of a variable (income), whose class distribution does not show a clear distinction between each class. Thus, this variable might not be as effective in discriminating high-risk vs. low-risk patients.

Figure 4: Distribution of median income based on readmission status

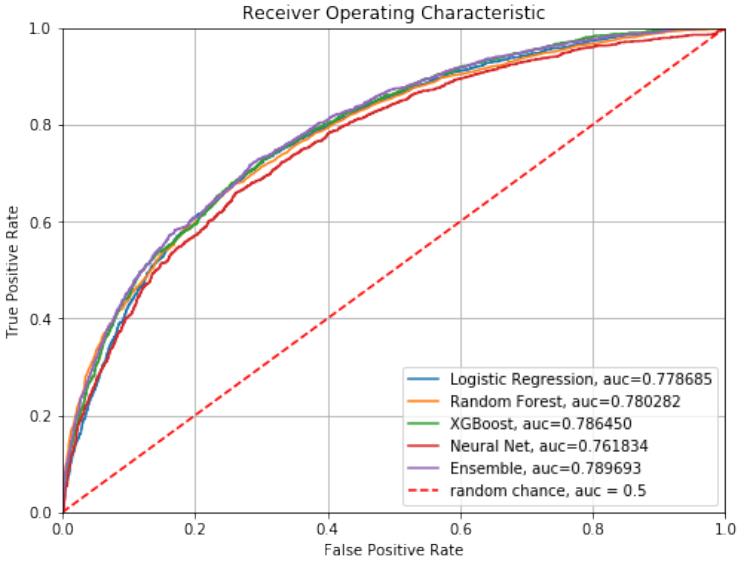


3.2 Model Performance

For our modeling process, we created four modeling criteria to evaluate the performance of our predictive models: interpretability, generalizability, discrimination, and calibration. We want an interpretable model because nurses and doctors will be the ones who utilize the results of our model. The third criterion, discrimination, refers to how well the model is able to classify patients correctly as being readmitted or not. The fourth criterion, calibration, refers to how accurately the model can produce risk scores that is closely aligned with the true probabilities for being readmitted. The last two metrics will be particularly important from a business standpoint because they both contribute to cost savings. This is because there is a cost associated with making type I errors (predicting that patients will be readmitted when they are not) or type II errors (predicting that patients will not be readmitted when they are). Since we want to create a model that is generalizable to the whole population, we partitioned our data into training and test sets (30%, 70%) so that we can test our model against unseen data. We utilized the training set to estimate population parameters and the test set to evaluate our model. We began with low-complexity models and gradually increased the complexity so as to maintain the interpretability aspect. For each model, we used the training data to perform cross validation for parameter tuning and evaluated our model using classification accuracy rate and AUC curves to ensure that our model has high

discrimination and calibration. We also incorporated regularization to ensure that we do not overfit the data. We fitted logistic regression with L1 regularization and found that the model yields an improvement from the baseline classification by 4 percentage points (from 73% to 77%), suggesting that our classifier performs better than the baseline. To achieve an even higher accuracy and AUC, we also tried non-linear and ensemble classifiers such as tree-based models (random forests and XGboost) and neural networks. Below is the result of each classification algorithm (figure 5). Apart from ensemble model, we can see that XGBoost performs the best and meets our criteria of generalizability and interpretability. Thus, we believe that this model will perform well in production and will integrate well with CommUnityCare's system.

Figure 5: AUC of ROC Curve for Classifiers



3.3 Local Interpretable Model-Agnostic Explanations (LIME)

In addition to our classification models, we also produced visualization tools for nurses and doctors to easily look up the risk profile of each patient. We incorporated a visualization technique called LIME, a technique invented by researchers from the University of Washington [7]. The algorithm uses linear models to approximate how changes in the value of input variables affect the predicted probabilities. This gives a sense of the relative importance of each risk factor for a given patient. Using this information, doctors and nurses can take

steps accordingly.

4 Recommendations and Business Value

4.1 Recommendations

Our classification model can predict a patient's probability of readmission at the time of admission. While a high predicted probability. While a high predicted probability indicates high readmission risk, it's hard to establish an optimal threshold for classification (i.e. when the predicted probability is higher than the predetermined threshold, classify the patient as high risk for readmission). We performed a cost sensitivity analysis to determine an appropriate threshold of a classifier to optimize cost savings, which we'll discuss in the next section.

Also, we believe that stratifying patients into high-risk and low-risk group alone is not effective/informative enough for nurses/doctors to to take appropriate preventative steps post-discharge. Hence, as mentioned in previous section, LIME was used to interpret the prediction from the classifiers. As shown below (figure 6 and figure 7), LIME outputs the important risk factors (features) for both positive ("Yes") and negative ("No") classes with relative importance. With the explanations from important risk factors for each individual encounter, nurses and doctors could have a more comprehensive understanding of the prediction from the classifier rather than simply treating it as a "black box". In this way, appropriate follow-up plans could be arranged accordingly for each patient based on their risk profile (i.e. through scheduling phone call, patient post-discharge education, home care or outpatient treatment).

Figure 6: LIME output for an encounter with high predicted probability for readmission

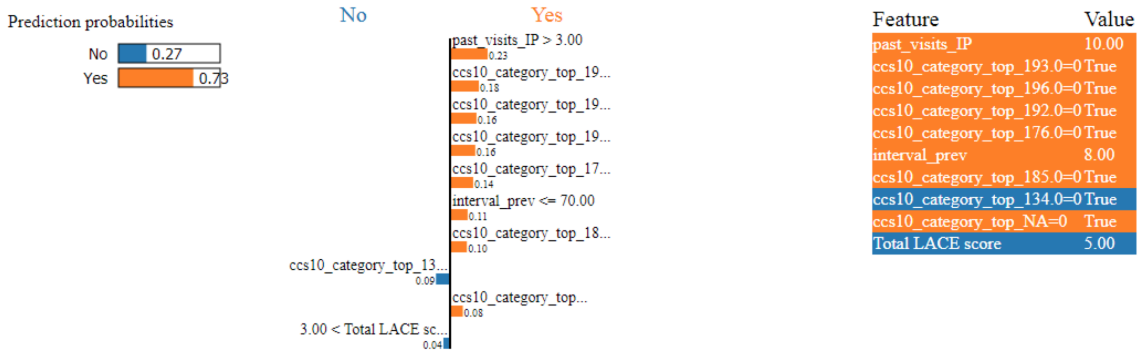
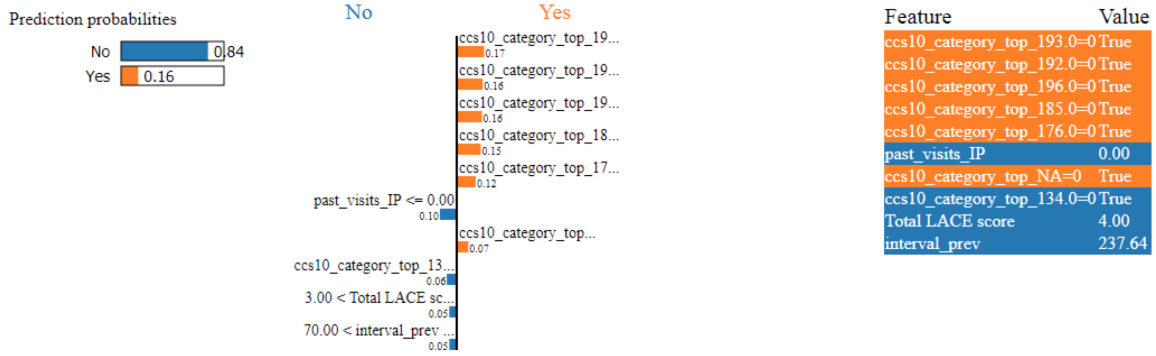


Figure 7: LIME output for an encounter with low predicted probability for readmission



4.2 Business Value

To obtain the cost-saving optimal threshold for the classifiers (e.g., for a threshold = 0.5, if the predicted probability = 0.7, then we should predict that the encounter will be readmitted), we performed a cost sensitivity analysis to quantify the financial impact of the classifiers. More specifically, we aimed to measure the monetary savings which could be potentially generated for CommUnityCare through the implementation of our predictive model.

We assumed that, by using our model, the care management team can take preventative steps post-discharge for for high risk patients (i.e. patients who predicted to be readmitted with the classifiers). We further assumed that performing these steps would result in a 20% reduction of readmission, based on the results from [8]. Hence, for each encounter predicted to be high risk (positive class), the cost of post-discharge cares β will be incurred. Out of all encounters predicted to be high risk, part of them won't be readmitted (i.e. a false

positive prediction); For the remaining correct predictions, there's a 20% probability that the post-discharge cares could effectively prevent such readmissions, hence resulting in savings of $0.2 \times$ average cost per readmission α.

Therefore, the savings matrix associated with identifying high risk diabetic patients for readmissions using the classifiers is defined as

$$Savings = \begin{bmatrix} \alpha - \beta & 0 \\ -\beta & 0 \end{bmatrix}$$

The objective of the classifiers is to maximize the expected savings, given the above savings matrix. Hence, appropriate threshold for the classifiers need to be determined.

The average cost of readmission for medicare, medicaid, privately-insured and uninsured patients is \$13,800, respectively [1]. Therefore, we calculated the weighted average of the readmission cost based on the distribution of payer from the dataset and estimated the readmission cost per patient of \$13,179. Adjusting for the effectiveness of post-discharge care, average cost savings per identified high risk patient α is $0.2 \times \$13,179 = \$2,635.80$. To calculate the average cost for post-discharge care, we used estimation from [9] and adjusted the cost to 90-day scale. The estimated cost β for post-discharge care is \$600 per high risk patient.

We then proceeded with finding the threshold denoted by $p(C_1 \mid x)$ (i.e., the probability of a given patient being classified as readmission) that maximizes the cost saving matrix by solving the equation below:

$$2635.8 \times p(C_1 \mid x) + (-600) \times (1 - p(C_1 \mid x)) = 0 \times p(C_1 \mid x) + 0 \times (1 - p(C_1 \mid x))$$

The solution to the equation above implies that, in order to maximize savings, we should classify a patient as readmission when the probabilistic prediction made by the model is above 0.1854.

Lastly, we calculated the potential cost savings which could be generated by each model as seen in the table below (table 1). Models with high discriminatory power and high calibration yield high savings. Furthermore, the value of the cost-saving optimal threshold indicates that the cost of post-discharge care is much cheaper comparing to cost of readmission. As shown, estimated cost savings per one thousand encounters (with 26% readmission rate) is \$0.386

million. If the probability of a post-discharge intervention preventing a readmission can be improved, the financial impact of this model would be even greater.

Table 1: Comparing estimated savings on hospital readmissions with different classifiers

Model	**Savings**	**Savings per 1000 encounters**
Ensemble	$2,567,952	$386,100
XGBoost	$2,525,092	$379,656
Random Forest	$2,519,802	$378,861
Logistic Regression	$2,503,131	$376,354

4.3 Limitations

Due to the quality of the dataset provided by CommUnityCare, several assumptions were made during the data pre-processing and feature engineering phases (e.g., imputing missing values by mean, dropping duplicated records, preserving most frequent CCS category for diagnosis). These assumptions were made based on our experience and best judgment, which will not necessarily lead to the best model performance. Also, readmission risk is greatly impacted by socio-economic factors. Although we utilized zip code to incorporate median income, education level and poverty rate information, geography-based information is not necessarily the best representation of an individual's socio-economic background. Lastly, a large proportion of readmissions are unpreventable. Although we set several criteria to filter out non-preventable readmissions, it s likely that a significant part of the readmissions used for modeling are non-preventable due to insufficient information on scheduled admissions, which could affect our model's performance in identifying preventable readmissions and could potentially cause ineffective resource allocation.

To optimize the potential cost savings resulting from risk stratification and related transitional care to prevent future readmissions, patient risk profiles constructed from classification models will be based on estimations of average costs of hospital readmissions and transitional care. Since our data does not include financial information for such costs, we used available industry data to produce these estimations. However, the actual costs could be different from our estimations which may lead to predictions that are not optimal for cost savings.

5 Conclusions

In conclusion, while we performed a thorough analysis to identify important risk factors and to estimate cost savings, the main difficulty lies in the limitations of our dataset as well as our limited domain knowledge on the industry. We wish that we could obtain more socio-economic factors on a higher level of granularity. We believe that such information can improve the predictive power of our model and will allow us to provide better recommendations. Future efforts could go into expanding the feature space of CommUnityCare's data to incorporate more personalized information about patients. Nevertheless, significant business value could be generated from incorporating predictive models to identify high risk patients. Our team strongly recommend CommUnityCare to devote more research effort in this area which will lead to enormous cost reduction and better healthcare outcome.

6 References

[1] K. Fingar and R. Washington, "Trends in hospital readmissions for four high-volume conditions, 2009–2013: statistical brief# 196," 2006.

[2] "Readmissions-reduction-program," Mar 2018.

[3] C. Boccuti and G. Casillas, "Aiming for fewer hospital u-turns: the medicare hospital readmission reduction program," *The Henry J. Kaiser Family Foundation*, pp. 1–10, 2015.

[4] "Readmissions medicare: Whats the cost?," Mar 2016.

[5] C. M. Ashton, D. J. Del Junco, J. Souchek, N. P. Wray, and C. L. Mansyur, "The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence," *Medical care*, vol. 35, no. 10, pp. 1044–1059, 1997.

[6] Y. Lax, M. Martinez, and N. M. Brown, "Social determinants of health and hospital readmission," *Pediatrics*, vol. 140, no. 5, p. e20171427, 2017.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, ACM, 2016.

[8] C. Jackson, M. Shahsahebi, T. Wedlake, and C. A. DuBard, "Timeliness of outpatient follow-up: an evidence-based approach for planning after hospital discharge," *The Annals of Family Medicine*, vol. 13, no. 2, pp. 115–122, 2015.

[9] J. S. Richardson, T. L. Mark, and R. McKeon, "The return on investment of postdischarge follow-up calls for suicidal ideation or deliberate self-harm," *Psychiatric services*, vol. 65, no. 8, pp. 1012–1019, 2014.

7 Acknowledgements

We would like to acknowledge our sponsors at CommUnityCare - Dr. Chris Pate, Dr. Eda Baykal-Caglar, Claire Dadic, and Dominic Armstrong - as well as our sponsors at Dell Medical School - Dr. Steven Andrews, Henry Robertson, and Dr. Anjum Khurshid - for their guidance and contribution of knowledge to this project. Their efforts in hosting the data, setting up a working environment, attending bi-weekly meetings, and consistent feedback was invaluable to the success of our final product. Additionally, we would like to thank our course instructors Dr. Michael Hasler and Dr. Ramesh Rajagopalan for arranging a capstone project between the McCombs School of Business MSBA program and CommUnityCare and Dell Medical School.