Group C: Zongyang Mou, Mary Rieger, Zixi Wang
MED 264 Final Project Report
Fall 2020 Quarter

Title: Improving prediction of adult ICU length of stay with Medication Administration Data

Abstract

Objective: Accurately predicting intensive care unit (ICU) length of stay (LOS) is important for patient care and stewardship of healthcare resources. However, current prediction models are limited. The objective of this project was to determine if augmenting a machine learning model with the number of medications administered within the first 24 hours of an ICU stay would offer improved prediction of patients' LOS in ICU.

Methods: We used custom PostgreSQL and Google BigQuery adapted from previous publicly available machine learning model studies to obtain adult admission data from the first 24 hours of ICU admission from the MIMIC-III database to build our patient population. This data was then fed into Random Forest Classifier, Decision Tree Classifier, Gaussian Naive Bayes, Stochastic Gradient Descent, Support Vector Machine with linear kernel, and Neural Network machine learning models. Cross-validation was then carried out to identify the models with the best performance and used then to create a combined model based on a voting algorithm. This combined model was then developed using the first 24 hour ICU stay data with medication administration vectors and compared against an identical model trained without the medication data, and feature importance was evaluated along with model metrics.

Results: Our study cohort consisted of 45209 unique ICU stays of adults admitted to various ICUs with an average LOS of 4 days. We found that a combination model of RCF, DTC, SVM, and NN created the best result through cross-validation with the single model RCF as offering the best AUC-ROC of 0.90. The combination model had an AUC-ROC overall of 0.91147 when the medication data is incorporated when compared with an AUC-ROC of 0.90648 in a model trained on data without the medication vectors. The highest performance gain with medication administration data was realized in the 3-7 days of ICU stay group with an F1 score of 0.7835 vs 0.7310. Overall medication administration was noted to be one of the most important features in our machine learning models.

Conclusion:
Medication administration is an important dimension that should be considered for incorporation in future improving machine learning efforts for predicting ICU LOS and should be optimized in the context of novel or existing models.

1 Introduction

The importance of good stewardship of hospital resources has been brought into the spotlight this year because of the COVID pandemic. Hospital intensive care unit

(ICU) capacity has been of particular concern. The ability to accurately predict a patient's ICU length of stay (LOS) is critical for hospital administration, who must continually assess allocation of hospital resources such as modifying elective surgery scheduling for patients who may require a postoperative ICU admission if the ICU is already at capacity with patients expected to have a long ICU stay. Additionally, ICU LOS prediction is valuable for a patient's family in making preparations for their family member's treatment.

Multiple efforts have been made to develop prediction models for ICU LOS, although there is not a consensus regarding which model is superior and the current ability to predict LOS is still limited.[1] A systematic review of prediction models for ICU LOS after coronary artery bypass graft surgery identified 12 papers on the development of new models all of which used linear or logistic regression, but the authors concluded that there was still need for methodological improvement, specifically with regard to standardizing the outcome and LOS risk factor definitions.[1] Widyastuti et al developed a preoperative and intraoperative prediction model for ICU LOS after cardiac surgery and also compared their models to existing models, but concluded that none of the models were useful at predicting LOS at the individual patient level.[2] Ghorbani et al evaluated the use of the Acute Physiology and Chronic Health Evaluation-IV score in their ICU patients and found it was a poor predictor of ICU course, significantly underestimating the LOS.[3] Brandi et al created a prediction model for pediatric ICU LOS, however they concluded that their model was not accurate enough to be adopted for discharge planning purposes.[4] Interestingly, another model aiming to predict ICU mortality used natural language processing (NLP) from ICU admission notes within the first 24 hours and found that, among others, the terms "intubated", "chemotherapy", and "sepsis" were significant predictors of mortality and terms "diet" and "awake" were predictors of survival.[5] The NLP based model for ICU LOS prediction is different from other models in that rather than using objective data as previous models, this one is a reflection of the gestalt of the treating physician with regards to how sick the patient is and thereby corresponds to a predicted LOS. This then prompted us to look for other objective measurements that are reflections of a treating physician's decision that may augment existing models. One evident measure is the number of medications administered during a patient's stay. Indeed there have been retrospective data that points to polypharmacy in hospitals as correlation to increased LOS, although this body of literature has not been tailored to the ICU population.[6] This is in line with our own clinical observations in which more acute ICU patients require more medications as interventions or surgeries than the less acutely ill patients. Thus, we hypothesized that augmenting a machine learning model with the number of medications administered within the first 24 hours of an ICU stay would offer an improved prediction of patients' LOSy= in ICU.

2 Materials and Methods

2.1 Data Acquisition and Patient Selection

We used the Medical Information Mart for Intensive Care III (MIMIC-III) critical care database to build our model.[7] The MIMIC-III database is comprised of the deidentified hospital records of more than 40,000 patients admitted to the Beth Israel Deaconess Medical Care critical care units between 2001 and 2012. This freely available

dataset is ideal to answer our study question because it specifically contains medical records data from a very large sample of ICU patients at a high level of detail. The large number of patients in the sample also lends well to machine learning methods and increases its generalizability.

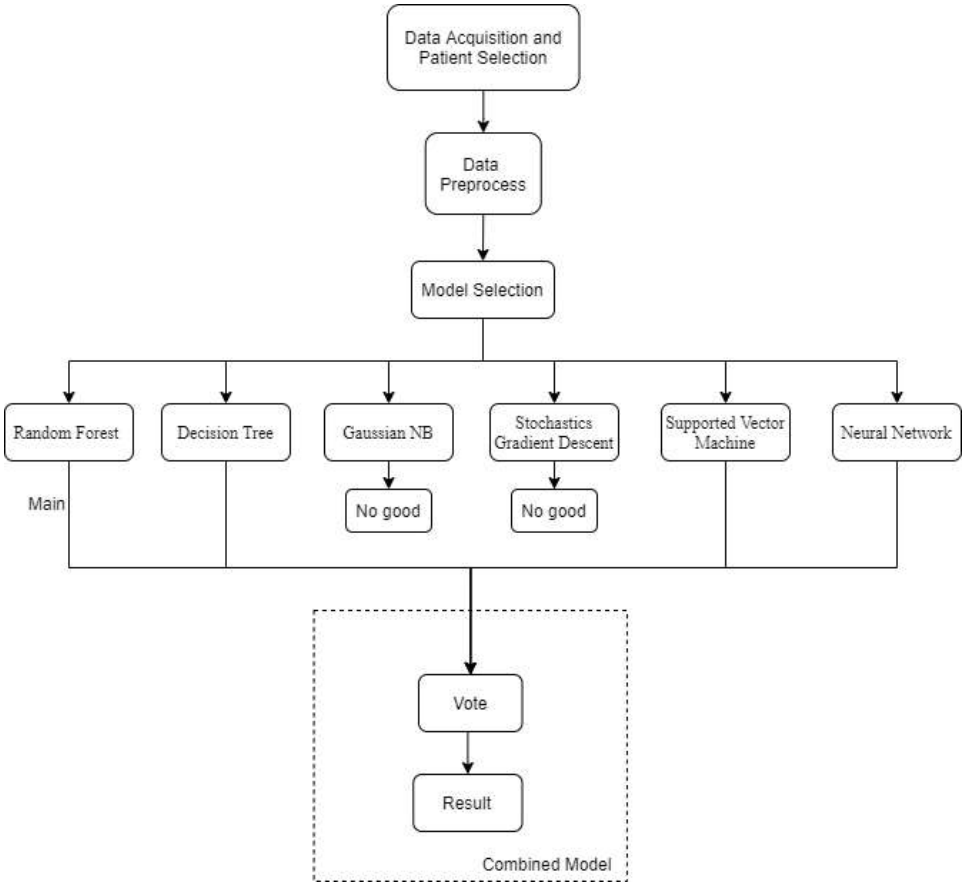
The main utility of this model is to project a LOS for resource management planning purposes, and this information would be best useful within the first 24-48 hours of an ICU admission. As such, we measured the total number of medications within the first 24 hours of an ICU as this achieved a balance of having enough information but early enough within an ICU stay to provide useful real-time information in clinical practice. Furthermore, we focused on the number of drugs given by continuous infusion and or by intravenous injection as a particular subcategory of interest because those are only given when patients are critically ill and are one of the criteria for hospital admission. We had initially also considered including specific medication types since this seemed to be a more direct surrogate for type and severity of illness. However upon inspection of the MIMIC-III dataset, we realized there were far too many different medications and any classification attempts would be prone to bias and require a separate project for validation to prevent misclassification errors. For instance, a drug such as metoprolol could be considered a vasoactive agent as well as an antiarrhythmic agent. In addition, the same drug can be used for different indications and circumstances and may not always reflect critical illness. As such, the nature of the actual medication was not incorporated within our model to reduce complexity.

To build our patient population, we used custom PostgreSQL and Google BigQuery adapted from previous machine learning model studies (code publically available on Github at https://github.com/illidanlab/urgent-care-comparative) to obtain the first 24 hour medication data, demographics, and other summary statistics related to that ICU admission such as the number of notes written or the number of lab results in that time frame.[8] We included patients admitted to all ICU except for the neonatal ICU such that our patient population would only contain adults above the age of 18 in our cohort. We did not select for a specific ICU unit knowing that different units will have different LOS, but we incorporated that into our machine learning model as a feature to account for this expected difference. For patients who had multiple ICU stays in a single admission, we used their first ICU stay as further ICU admissions within a hospital admission can be confounded by other factors and have inaccurate drug data. Patients who had multiple unique admissions were included in our data set and treated as independent samples.

2.2 Machine Learning Models

We created our machine learning classification model based on the ideas in *Predictive Modeling in Urgent Care: A Comparative Study of Machine Learning Approaches* [8] and *Predict Hospital Length of Stay - Classification* [9]. The goal of the model is to predict the ICU LOS based on the given data, and show the significance of the data collected in the first 24 hours of the ICU stay. The workflow diagram is shown in Figure 1.

Figure 1. Workflow diagram



2.2.1 Preprocessing

The LOS of each patient in the dataset is a float value such as "1.3 days". With the consideration of the practicality and the performance of the model, we grouped the LOS into 4 classes: 0 - 1.5 days, 1.5 - 3 days, 3 - 7 days, and more than 7 days and used these as labels of the data. To the physicians in our group, these classifications would represent meaningfully different categories of LOS for both hospital administration and patient family member planning purposes. We did not think that a more specific prediction, such as a specific number of hours of LOS would be necessary or provide clinically meaningful information. Then we dropped marital status, ethnicity, religion, and insurance information to make the result less likely to have a negative bias against any social group. The text data had to be converted to numerical data in order to train the model. We used one-hot encoding to make each unique text value a feature with values 0 and 1, indicating whether the datapoint has the text value or not. The last step was to split the data into a 20% test set and an 80% train set with 4 folds for cross-validation.

2.2.2 Model Training

In order to find out which models are effective, we experimented with the Random Forest Classifier (RFC), Decision Tree Classifier (DTC), Gaussian Naive Bayes (Gaussian

NB), Stochastic Gradient Descent (SGD), Support Vector Machine (SVM) with linear kernel, and Neural Network (NN) with 5 dense layers (relu activation) and 5 dropout layers (0.5 dropout rate). For each model except NN, we performed grad searches to optimize the hyperparameters. The grid search for NN took an unacceptably long time to run so we adjusted the hyper-parameters manually. For each model, we performed 4-fold cross-validation to prevent overfitting and plotted the receiver operating characteristic curve to estimate the performance of the model. We then selected the models with good performance to form a combined model, which is defined as having a cross-validation score and AUC > 0.8. For the combined model, the result is voted by each model. If there is a tie, the result follows the vote of the model with the best cross-validation score. Then, we generated the confusion matrix for each class to analyze the results. Finally, we generated a chart from the model with the best performance to show the importance of significant clinical features. We used two sets of data to train the model for comparison: the dataset contains medication data collected in the first 24 hours of the ICU stay and the dataset without the first 24 hours of medication data.

2.3 Statistical Analysis

The statistical software R was used for generating descriptive statistics for our patient cohort and comparing across categorical levels. 2-sample independent T-test was used to compare the LOS between males and females. 1-way ANOVA was used to examine the LOS across ICU units with an overall threshold P-value of 0.05 and pairwise comparison adjustment for multiple comparisons as defined by 0.05/10 for 10 pairwise comparisons. Built-in model metrics for the machine learning models including accuracy, recall, precision, F1 Score, area under the curve of receiver operator curves (AUC-ROC) were generated using scikit-learn library for Python.

3 Results

3.1 Study Cohort Description

We identified 45,209 unique ICU admissions in our query cohort. There was a slightly higher percentage of males than females (56.2% vs 43.8%, Table 1). The average age of the patient was 61.5 years old with a standard deviation of 16.3 years (Table 2). The majority of the patients were of Caucasian descent at 73%, and the majority of patients had private insurance at 31% (data not shown). The majority of patients (40%) were admitted to the medical ICU (MICU) followed by surgical (SICU) and cardiovascular surgical ICU (CSRU) at 16.5% each (Table 1). In the first 24 hours, each patient received on average 16 total medications, 2 continuous IV medications, and 7 IV injection medications (Table 2). The average LOS for each ICU stay was 4.1 days with a SD of 6 days (Table 2). When stratified by gender, there was no difference between men and women (P=0.318, Table 1). As previously demonstrated, the unit type is significantly associated with LOS with those admitted to the SICU and trauma ICU have a longer length of stay by about half a day on average compared to other ICU units (overall P-value <0.001, Table 1).

Table 1. Distribution of cases by select categorical variables with mean (SD) of LOS by variables. 2-sample independent t-test or 1-way ANOVA were used for overall tests of significance. Significant pairwise comparisons against the top-level reference group shown with P-value of <0.05 considered significant.

Variable	N (%) Total N = 45209	Mean (SD) LOS in days	P value
Gender			**0.318**
Male	25,416 (56.2%)	4.1 (6.1)	
Female	19,793 (43.8%)	4.0 (5.8)	
ICU Unit Type			**<0.001**
CCU	6390 (14.1%)	3.9 (5.6)	
CSRU	7449 (16.5%)	3.8 (5.8)	0.381
MICU	18191 (40.2%)	3.9 (5.7)	0.440
SICU	7478 (16.5%)	4.6 (6.8)	**<0.001**
TSICU	5701 (12.6%)	4.4 (6.3)	**<0.001**

2-sample independent t-test or 1-way ANOVA were used for overall tests of significance. Significant pairwise comparisons against the top level reference group shown with P-value of <0.05 considered significant. CCU, coronary care unit; CSRU, cardiac surgery recovery unit; MICU, medical intensive care unit; SICU, surgical intensive care unit; TSICU, trauma/surgical intensive care unit.

Table 2. Summary of values in the first 24 hours of ICU admission for continuous variables for the patients used in the model.

Variable	Mean (SD)
Age (years)	61.5 (16.3)
Length of stay (days)	4.1 (6.0)
Number of chart events	585.1 (640.1)
Number of CPT events	1.3 (1.3)
Number of diagnoses	2.6 (6.0)
Number of IV drugs (drips)*	1.6 (2.5)

Number of IV drugs (injections)**	6.5 (5.5)
Number of drugs (total)	15.5 (10.6)
Number of labs	50.5 (56.5)
Number of microbiology labs	1.3 (3.6)
Number of notes	2.5 (43.7)
Number of procedures	0.7 (2.2)
Number of unit transfers	0.9 (1.3)

*Drips refers to medications administered as continuous intravenous infusions.
**Injections refers to medications given as an one-time intravenous administration.
IV, intravenous; CPT, current procedural terminology.

3.2 Model Selection

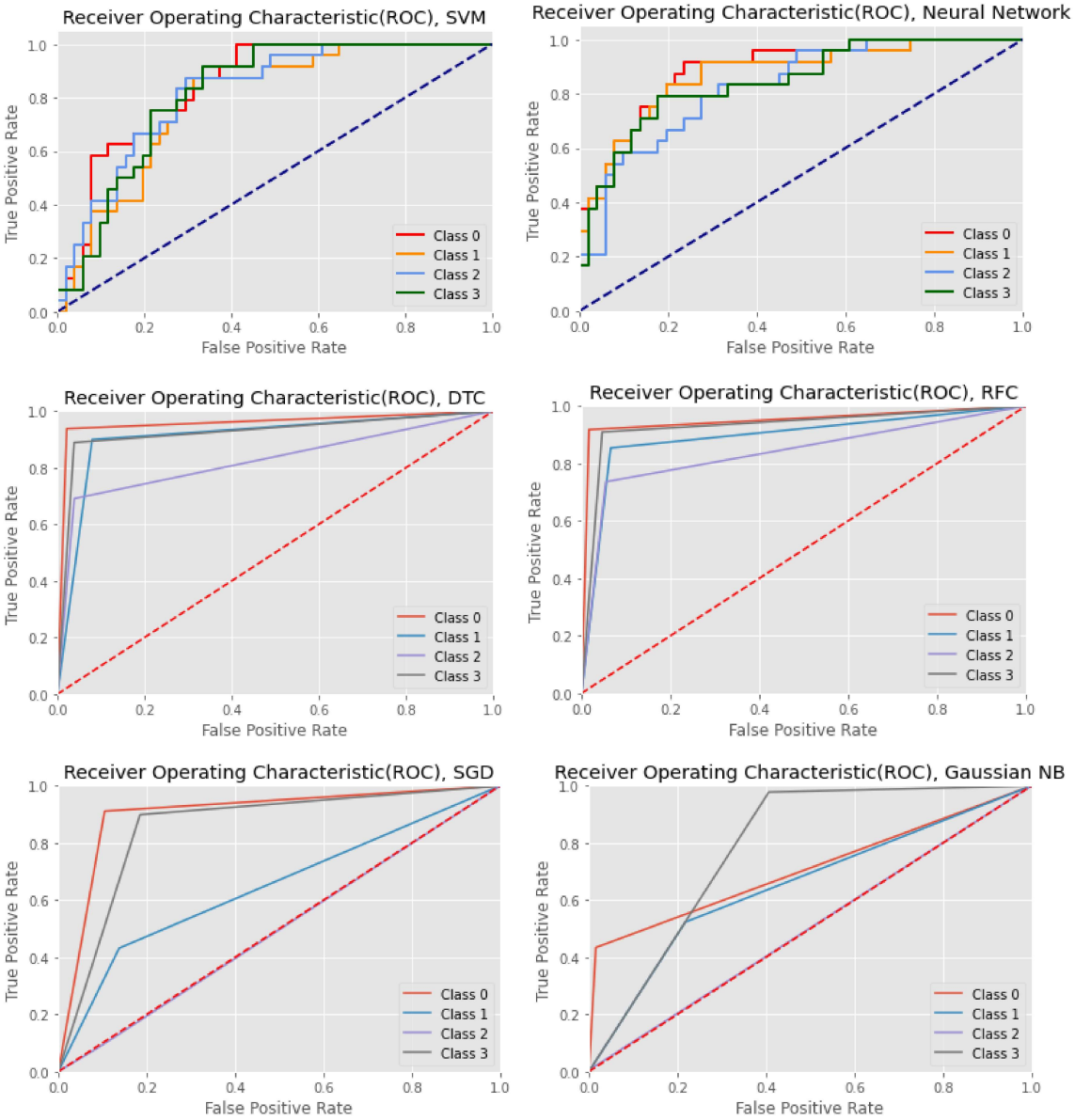
In order to screen the less effective models, we evaluated each model by its cross-validation score and AUC-ROC after the grid search (Table 3). The ROC is the receiver operating characteristic curve, which demonstrates the true positive rate compared with the false positive rate. The area under the ROC curve is called the AUC-ROC, which is a popular metric for classification performance. Our analysis showed that RFC, DTC, SVM, and NN were the best models to be used in our combination model and Gaussian NB and SGD were thus excluded accordingly. We generated the ROC curves for each model trained with the first 24 hours ICU dataset are shown in Figure 2.

Table 3. Performance table for models trained with dataset *without* the first 24 hours ICU data

Model	**Cross-validation score**	**AUC-ROC**	**In the combined model**
RFC	0.858213	0.900156	Included, Main model
DTC	0.845473	0.899745	Included
Gaussian NB	0.429835	0.546214	Excluded
SGD	0.638409	0.737237	Excluded
SVM	0.812301	0.897471	Included
NN	0.808647	0.894365	Included

AUC-ROC, area under the receiver operating characteristic curve; RFC, Random Forest Classifier; DTC, Decision Tree Classifier; NB, Naive Bayes; SGD, Stochastic Gradient Descent; SVM, Support Vector Machine; NN, Neural Network.

Figure 2. ROC curves for each model, trained with dataset WITH the first 24 hours ICU data



*Class 0 = 0-1.5 days stay, class 1 = 1.5-3 days, class 2= 3-7 days, class 3 = >7 days

3.3 Combined Model Performance

With RFC being the main model, the combined model contains RFC, DTC, SVM, and NN. The performance of the combined classification model is described in Table 4 and Table 5. In Tables 4 and 5, Accuracy refers to how often the classifier is correct overall. Recall indicates how often the classifier is correct for cases whose ground truth is positive (true positive rate). Precision describes how often the classifier is correct when the prediction is positive. F1 Score considers both Recall and Precision, emphasizing the effect of false positives and false negatives. As demonstrated by Tables 4 and 5, after

considering the medication data collected in the first 24 hours of ICU stay, the performance of the model increased for every class in each metric. There was a slight increase in AUC-ROC from 0.90648 to 0.91147, net increase of 0.005. When broken down by class, the highest absolute F1 Score gain was noted in the class 2 group with an increase from 0.7310 to 0.7835 (net gain of 0.0525) which was largely influenced by the improved Recall rate (0.7499 from 0.6734). Conversely, the class with the least improvement unsurprisingly is class 0 or those with <1.5 days within the ICU (net gain of 0.0093), although this group had the highest F1 Score across all classes in both models.

Table 4. Confusion matrix for model trained with dataset WITHOUT the first 24 hours ICU data (Multiclass AUC-ROC = 0.90648)

Length of Stay	Accuracy	Recall	Precision	F1 Score
Class 0 (0 - 1.5 days)	0.9705	0.9442	0.9279	0.9360
Class 1 (1.5 - 3 days)	0.9166	0.8999	0.8443	0.8712
Class 2 (3 - 7 days)	0.9140	0.6734	0.7994	0.7310
Class 3 (7 - 90 days)	0.9449	0.9077	0.8992	0.9035

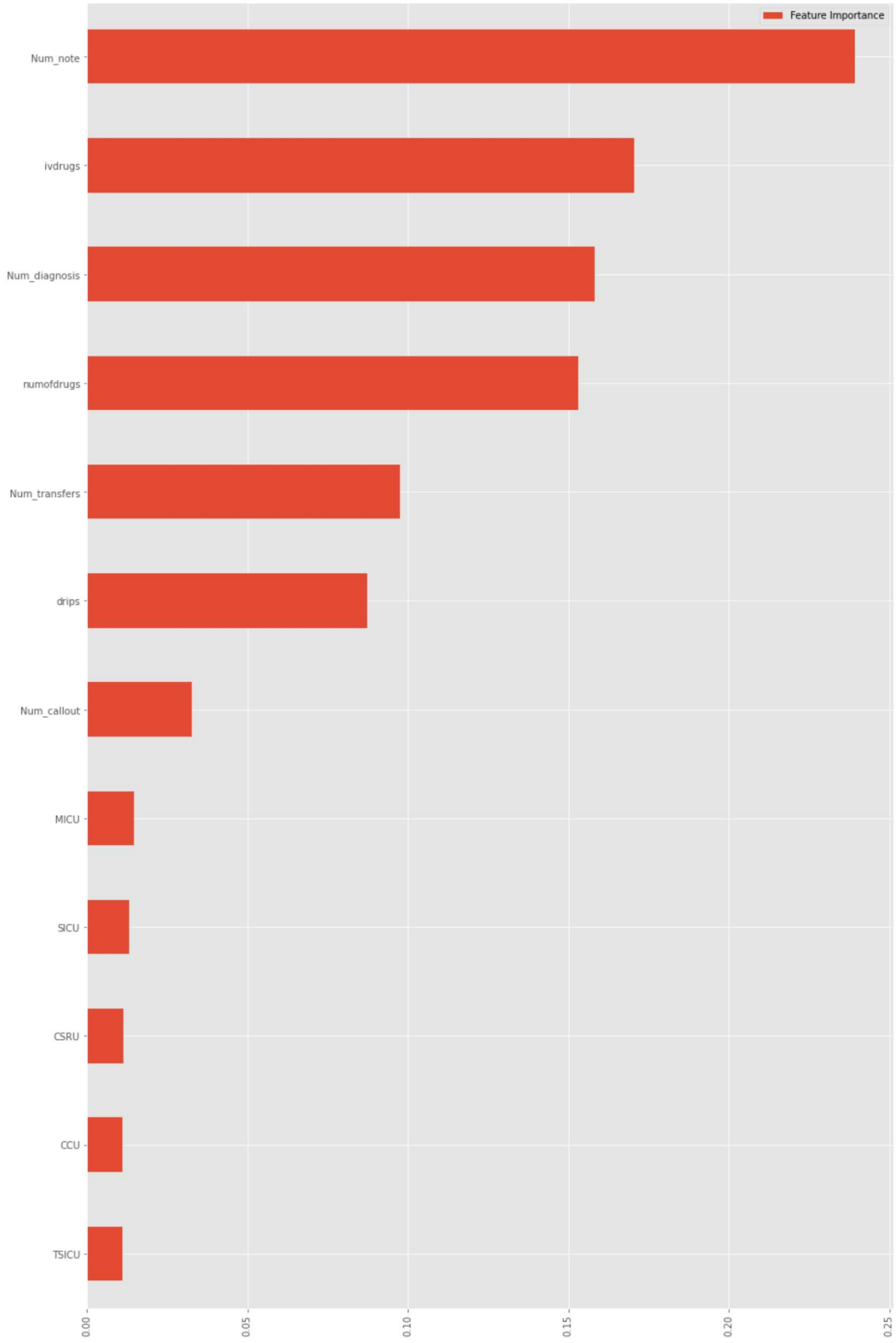
Table 5. Confusion matrix for model trained with dataset WITH the first 24 hours ICU data (Multiclass AUC-ROC = 0.91147)

Length of Stay	Accuracy	Recall	Precision	F1 Score
Class 0 (0 - 1.5 days)	0.9749	0.9492	0.9415	0.9453
Class 1 (1.5 - 3 days)	0.9335	0.9206	0.8741	0.8968
Class 2 (3 - 7 days)	0.9281	0.7499	0.8204	0.7835
Class 3 (7 - 90 days)	0.9530	0.9110	0.9227	0.9168

3.4 Feature Importance

In order to investigate the features that have significant effects on the result, the feature importance of the top significant features was generated based on the single model with the best performance, the RFC with the dataset containing the first 24 hours ICU data, using the model.feature_importances function in the scikit-learn library. The number of clinical notes, the number of IV drugs, the number of diagnoses, and the total number of drugs appeared to be of particularly high importance to predict ICU LOS (Figure 3).

Figure 3. Feature Importance



4 Discussion

Prediction of ICU LOS has been a challenge using prior ML models that have largely focused on readily accessible objective vitals of a patient's physiological status through lab values and vital signs and clinical exams. Here, we created a model that focused on leveraging the medication administration data to improve predictive performance. We found that using the number of medications within the first 24 hours had an incremental improvement in the overall performance of prediction of ICU LOS as demonstrated by the improved AUC score when compared to a model that did not use the medication administration data. Our comparison model matches in AUC roughly with previous models that demonstrated performances of AUCs ranging from 0.5 to 0.9.[8] Furthermore, we demonstrated that the number of medications in the first 24 hours is a feature of higher significance than other measurements such as the number of labs drawn for a patient. Of note, when comparing F1 scores, our model had the most significant absolute value gain in those who have a LOS of 3-7 days. This is of particular significance as this is within the time frame for which ICU resource management prediction is particularly useful as nursing and staffing shifts are arranged based on hospital bed projected needs in the coming days to week. Thus, having a better idea of the ICU bed availability in this time frame is particularly important for this purpose. Interestingly, we also found that the type of care unit for ICU admission was an important predictor. This is consistent with previous findings that ICU LOS is likely largely influenced by discharge policies distinct to each type of care unit.[2] A yet unresolved question then is how to incorporate these institutional policies changes into the models or to strive for more uniform discharge policies as different hospitals will have differing ICU protocols.

Our study was not without limitations, the most significant of which was the lack of more complex feature selections including trends of first 24 hour vital signs and laboratory values as we were not able to adequately represent the complexity of this data with the confines of our resources. However, we were able to demonstrate that even with our 13 feature model, we had comparable predictive values at baseline without the medication administration data. Another limitation is as stated before that this is a single institution database and so policies specific to an ICU and hospital limit the generalizability of this specific model to others as there is a significant interaction between those policies and LOS in an ICU. One other important limitation is that the prediction models still require generous grouping of LOS to multi-day bins rather than daily bins which limits its applicability. Lastly, there is noted relatively small incremental gain in performance metrics with the addition of the medication features to our baseline model. This may be explained by the fact that our model did not have sufficient complexity (in part due to lack of available computational resources) to model interactions between this and significant features such as vital signs that could increase the predictive value of the model.

With the fact that modeling a patient's physiological status based on objective values has not been the panacea for accurate ICU LOS prediction by machine learning models, it is imperative to look for additional dimensions of data to offer stepwise performance improvements. Previous research demonstrated the importance of this with NLP methods that represent the clinician's assessment of the patient's status in their notes as a valuable added dimension. Our work suggests that medication administration is

another dimension that should be considered for incorporation in future improving machine learning efforts for predicting ICU LOS and this feature should be further optimized in the context of novel or existing models.

5 Conclusion

In conclusion, we found that the incorporation of the first 24 hours of medication administration in an ICU was predictive of the LOS in the ICU and inclusion in machine learning models improved the performance of the overall AUC. Our model in particular demonstrated noticeable improvement in predicting LOS in those who stay in the ICU for 4-7 days. Our results suggest that future machine learning models looking at predictive LOS in the ICU should include the medication administration in the first 24 hours as a feature and could realize further performance gains by incorporating this feature.

6 References

1. Atashi A, Verburg IW, Karim H, et al. Models to predict length of stay in the intensive Care unit after coronary artery bypass grafting: A systematic review. *J Cardiovasc Surg (Torino)*. 2018. doi:10.23736/s0021-9509.18.09847-6
2. Widyastuti Y, Stenseth R, Wahba A, Pleym H, Videm V. Length of intensive care unit stay following cardiac surgery: Is it impossible to find a universal prediction model? *Interact Cardiovasc Thorac Surg*. 2012. doi:10.1093/icvts/ivs302
3. Ghorbani M, Ghaem H, Rezaianzadeh A, Shayan Z, Zand F, Nikandish R. A study on the efficacy of APACHE-IV for predicting mortality and length of stay in an intensive care unit in Iran. *F1000Research*. 2017. doi:10.12688/f1000research.12290.1
4. Brandi S, Troster EJ, Cunha ML da R. Length of stay in pediatric intensive care unit: prediction model. *Einstein (Sao Paulo)*. 2020;18:eAO5476. doi:10.31744/einstein_journal/2020AO5476
5. Marafino BJ, Park M, Davies JM, et al. Validation of Prediction Models for Critical Care Outcomes Using Natural Language Processing of Electronic Health Record Data. *JAMA Netw open*. 2018.
6. Teo V, Toh MR, Kwan YH, Raaj S, Tan SY, Tan JZ. Association between Total Daily Doses with duration of hospitalization among readmitted patients in a multi-ethnic Asian population. *Saudi Pharm J*. 2015;23(4):388-396. doi:10.1016/j.jsps.2015.01.013
7. Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. *Sci Data*. 2016. doi:10.1038/sdata.2016.35
8. Tang, F., Xiao, C., Wang, F., & Zhou, J. (2018, June 04). Predictive modeling in urgent care: A comparative study of machine learning approaches. Retrieved December 09, 2020.
9. Drscarlat. (2019, December 27). Predict Hospital Length of Stay - Classification. Retrieved December 09, 2020, from https://www.kaggle.com/drscarlat/predict-hospital-length-of-stay-classification