In-Hospital Mortality Predictions using Topic Modeling in Apache Spark

Bryan Travis Smith

***Abstract*—In this paper, we modeled the probability of in-hospital mortality once a patient is admitted to a hospital. We combined the admission-time information such as age, gender, and admission types with severity scores and Latent Dirichlet Allocation topics from clinical notes to build a logistic model for predicting an in-hospital mortality event. This method is applied to the MIMIC III dataset with 46,520 patients and 5813 in-hospital mortality events using Apache Spark for data processing, model building, and scoring. This work validates previous work suggested features generated from Latent Dirichlet Allocation are effective. The resulting models are informative for rank ordering the severity of patients in a real-time context. Retrospective AUC on in-hospital mortality predictions is found to be between 0.869 - 0.909 in cross-validation, and batched-real-time predictions produce AUCs between 0.80 and 0.85. This work can be easily adapted to a real-time processing and scoring of streaming medical-records.**

***Index Terms*—Big data, Health analytics, Data mining, Machine learning, Spark, Topic Modeling**

I. Introduction

BIG data and health-care applications interact closely, thanks to the advancement in electronic data capturing technology such as electronic health records, on-body sensors and genome sequencing. Technologies such as Apache Kafka are capable of creating streams of electronic medical records that can be consumed by Apache Spark for real- time processing, aggregation, predictions, and reporting of patient information. This information can help augment medical practitioners judgments and understanding about patient conditions with the goal of improving medical care and outcomes. Because doctors in hospitals, particularly Intensive Care Units (ICUs), are responsible for administering care among multiple patients with a wide variety of diseases and severities, prioritizing patient care can become amibiguous. Predicting a patient's mortality risk can act as a surrogate for severity of a patients condition.

This work focuses on using Apache Spark to reproduce previous results demonstrating the effectiveness of topic modeling using Latent Dirichlet Allocation (LDA) models on patient notes for predicting in-hospital mortality events. [1] Many have built on the performance of these insights. Nozomi et. al. did simultaneous modeling of multiple diseases with AUC's as large as 0.795 [2], Ghassemi has included time-series modeling with AUC's as large as 0.812 [3], and Caballero et. al. used time-series with topics, expectation maximization, and kalman filters to produce AUC's as large as 0.866. [4]. The retrospective predictions of these works are similar to the original Ghassemi et. al. results.

The work in this papers differs from the previous works in the following ways:

1) The previous publications used the Medical Information Mart for Intensive Care (MIMIC) II dataset while this paper makes use of the MIMIC III dataset. [5]
2) This works makes predictions based on admission to the hospital instead of admission to an Intensive Care Unit (ICU).
3) No patients were excluded by age (i.e. newborns) or by thin files.
4) All code and data was built using Scala and Apache Spark for scalability.
5) The model used is logistic regression for ease of interpretability.

We would like to highlight the last point. If a patient has an increase or a decrease in severity predictions from one day to another, an explanation to the difference in the prediction should be provided. Probabilistic interpretation of SVM are translated as a distance from the decision boundary constructed in fitting the model. The boundary itself, however, is not directly interpretable outside of margin-metrics in feature-space. Logistic models, on the other hand, have probabilities directly based on odds ratios using patient features, and a decision boundary is then constructed from a threshold on the probabilities. We believe that directly explaining predictions and changes in predictions will lead to validation, stress-testing, trust, accountability, and ultimately the use of machine learning models in medical settings. Models, such as SVM, have the benefit of avoiding over-fitting when the number of features are larger than the number of observations, but this work does not face this problem.

II. Methods

The work by Ghassemi et. al. highlighted that retrospective analysis, while showing better performance, are not an accurate representation of practical model performance. They also highlight that there is a delay between the documenting of information, and the uploading of medical records are entered into the system. Practical model implementations must take this reality into account. We attempt to reproduce a time series score that accumulates more information about each patient as more records are generated from their stay in the hospital.

A. Problem Formulation

The problem we address in this paper is attempting to assign patients daily probabilities of having an in-hospital mortality

event from data available in electronic medical records. The longer each patient is in the hospital, the more information is generated for each patient to better assess the severity of their condition. Models that can look back and make accurate predictions are useful for determining which features are predictive, but are not necessarily useful in practice.

B. Approach and Implementation

We used the MIMIC III 1.3 dataset [5] which has 46,520 patients, 58976 hospital admissions, 2078705 clinical notes, and 5813 in-hospital mortality events. We extract age, gender, admission type, and mortality outcome from the admissions records. We then construct an Oxford Acute Severity of Illness Score (OASIS), the Simplified Acute Physiology Score (SAPS) severity score, and Sepsis-related Organ Failure Assessment (SOFA) severity score from the first 24 hours of data for each patient. Patients were divided into a training-test split of 70% of patients assigned to the training set, and 30% of patients assigned to the testing set. This allowed a patient's complete record over multiple hospital and ICU admission to be fully contained within one set.

We used the training set of patients, clinical notes and excluded discharge notes, notes without proper storage times, and notes with errors, to build a series of models including a Term Frequency Model, a TF-IDF model, and an LDA topic model. Test data was never used to build a vocabulary, determine which vocabulary was informative, or used to build topics. We began the process by using all notes for a hospital-admission and combined the notes after the removing words from the Onix stop word list as well as common units of lab results. Each patient's notes were tokenized and turned into a term frequency vector that was limited to a vocabulary of 500,000 and a limit that a given token must appear in at least 20 collections of notes. The term-frequency vector was transformed to a TF-IDF vector for each patient-hospital admission combination. These steps reduce the vocabulary to approximately 23,000 terms.

An LDA topic model of 35 topics with topic-distributions prior set to 1.0 and topic-word prior set to 0.01 was constructed using the TF-IDF vector for each patient/hospital-admission combination. A feature vector is then constructed using the patients age, gender, admission type, severity scores if existing, and topics, to build a logistic regression model predicting in-hospital mortality. The summary of the topics words and log-odds for each topic is shown in Appendix A.

III. Results

The models produced by the process outline in the previous section are evaluated by the Area Under the Receiver-Operator Curve (AUC), which can be interpreted as the quality of the rank ordering of a patient's likelihood of having a mortality event. A value of 1.0 would be a perfect rank-order of severity, while a value of 0.5 would be a random rank-ordering. The papers outline in the introduction had AUCs between 0.8 - 0.9 for time-varying predictions, and above 0.9 for retrospective predictions.

A. Experimental Design

We have two evaluations of our models on the test set. The first is to do a retrospective evaluation by creating a 24 hour prediction window for each admission, where data is removed from the system if it is acquired less than 24 hours of patient discharge. The data is then transformed into feature vectors through the models built on the training data, then scored for prediction. The AUC metric is then calculated for all admissions. It is worth noting that all patients that stay less than 24 hours in the hospital will only have the initial admission features.

The second way we evaluate our model is to evaluate it on a time-varying basis. We do this by collecting all the data with the evaluation time period, constructing the features for that admission and within that time period, and evaluate the prediction that the admission will result in an in-hospital mortality. There is some subtlety to this evaluation because each admission has a different length of stay. If a patient has a length of stay less than 48 hours, they are not evaluated in windows that are larger than 48 hours. Figure 1 illustrates the count of admissions remaining vs length of stay (top) and the percent admissions that turn into an in-hospital mortality events (bottom).

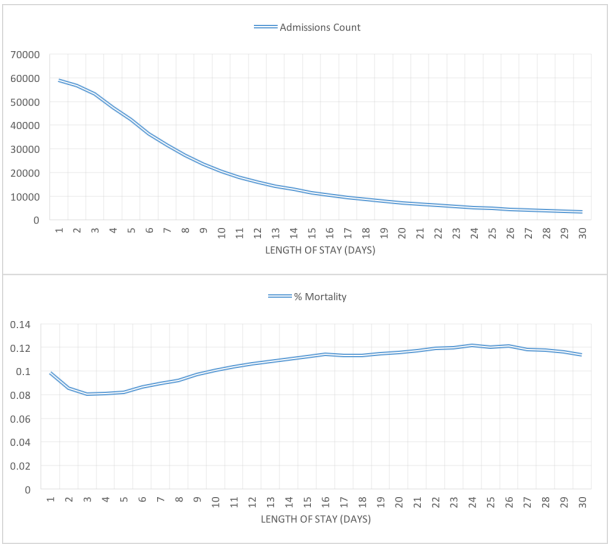

Fig. 1. Top: Total number of remaining admissions as a function of number of days in the hospital. Bottom: Percent of remaining admissions that result to an in-hospital mortality event as a function of the number of days in the hospital.

B. Experimental Results

The retrospective analysis of our process was done by rerunning the entire model building process on a 10-Fold cross validation where 70% of the data was used to rebuild the model and than score the AUC on the remaining 30%. The AUCs ranged between 0.869 and 0.909 with an average of 0.878.

The time-varying evaluation of the model is shown in Figure 2. The bottom (blue) line is just using the admission time as a baseline. This prediction is made from the first 24 hours of data, and is not updated at any point after this initial prediction.

The fact that the models ability discriminates the severity of patients as their time in the hospital increase is expected. The middle (orange) curve is a logistic model built only on the time varying topics. This model is much better than the admission-based model on discriminating the severity of patients. It is unexpected that the models ability to discriminate severity after 5 days degrades. More clinical notes add predictive power in the beginning, but do not continue over time. This suggests that a rolling window of notes might be better than a pure combination of all clinical notes. The top (gray) line is the final time-varying model that uses admission-based features as well as time varying topics. At a high level, the results of this analysis match the results from Ghassemi et. al. for the combined model predicting in-hospital model events.

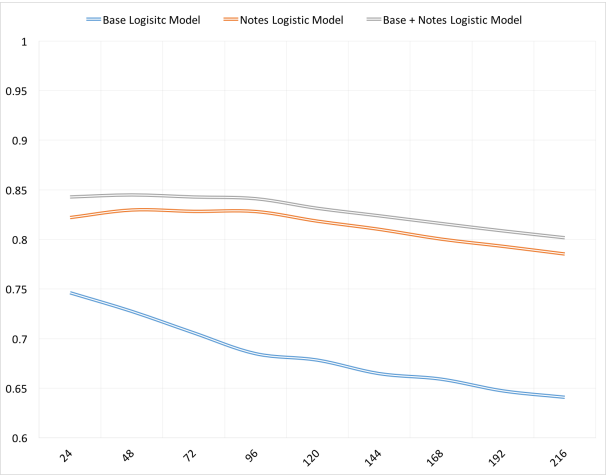

Fig. 2. Graph of AUC of model predictions on test data for time-varying features. The bottom (blue) line is the baseline admission model. The middle (orange) line is the performance of a model based only on topics. The top (gray) line is the performance of the final model using both admission data and note topics.

IV. Conclusions

This work was motivated by Ghassemi et. al.s discovery that there is rich and useful information in hospital notes that inform a given patient's severity and that time-varying models could potentially be useful in a hospital context. Real-time models require streaming context and a large amount of data, so reproducing their work in Scala with Apache Spark is a step towards scalable real-time medical predictions as well as a validation of the process of using clinical notes is informative in prediction patient outcomes.

Despite the increase rate of in-hospital mortality with longer hospital states, the MIMIC III dataset highlights that the window of time with the most mortalities is the first 24 hours. This windows makes up 16.7% of all mortalities in the dataset. The model in this paper is not capable of scoring these patients before they have died.

Another limitation of the current work can be seen in examining topic 5 of appendix A. It shows an exceptionally high odds-ratio for mortality in the model. The top 20 words for this topic is associated with automobile accidents. Despite being able to describe the severity of this event, and the LDA creating a topic for it, a model for the severity of this condition is likely not adding information that the medical practitioner in the hospital do not already have.

Topics 10 and 18 both have the words pt and goal. Each of these topics topics have odds ratios less than 1, and the words in these topics appear to be associated with recovery. The model describes them as associated with lower risk of mortality. The issue with this aspect of the process and model is that the information is in the system after a doctor has applied judgment to order physical therapy and a physical therapist has engaged the patient. In these cases, the model is reflecting back practitioners' judgment instead of augmenting it.

In both of the above examples, the model is accurately capturing in-hospital mortality risk, but in situations that are not possible to predict. The value-add for severity scores are in situations where the immediate risk is not obvious to an experienced medical practitioner. It seems that models built on clinical notes are capable of capturing and describing a patient's states and risks of in-hospital mortality. It also appears that some of the clinical notes capture the physician's judgment about the severity of the patient's condition, and the prediction system reflects this back after the notes have been processed.

There is more work to be done to determine the optimal type and range of features to be included in streaming medical reporting and prediction systems. The information and prediction should be informative and predictive to be both practical and useful. Clinical notes have been confirmed to be informative in our work. The subtlety in how they are used to generate predictive features still needs to be discovered.

References

[1] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State: Mortality Modeling in Intensive Care Units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 75–84, New York, NY, USA, 2014. ACM.

[2] Nozomi Nori , Hisashi Kashima , Kazuto Yamashita , Hiroshi Ikai , Yuichi Imanaka, Simultaneous Modeling of Multiple Diseases for Mortality Prediction in Acute Hospital Care, In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 10-13, 2015, Sydney, NSW, Australia

[3] Marzyeh Ghassemi , Marco A. F. Pimentel , Tristan Naumann , Thomas Brennan , David A. Clifton , Peter Szolovits , Mengling Feng, A multivariate time-series modeling approach to severity of illness assessment and forecasting in ICU with sparse, heterogeneous clinical data, In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, p.446-453, January 25-30, 2015, Austin, Texas

[4] Karla L. Caballero, Ram Akella, Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach, In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 10-13, 2015, Sydney, NSW, Australia

[5] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, May 2011.

APPENDIX A
TOP 20 WORDS AND ODD-RATIO FOR EACH TOPIC

Topic	Odds Ratio	Top 20 Words
1	31.9	fx, fracture, ortho, rib, hip, action, assessment, trauma, fractures, orif, femur, fusion, pain, response, injuries, medications, dilaudid, knee, brace, laceration
2	5.2	infant, nicu, sepsis, baby, maternal, cbc, gbs, parents, nursery, term, newborn, murmur, intrapartum, born, normal, nbn, pregnancy, warmer, delivery, fetal
3	2.3	dementia, mental, history, assessed, javascript, altered, icu, popup, meq, pulse, webtag, action, fever, medications, pneumonia, patient, assessment, myeloma, mri, alzheimer
4	0.56	lap, tpn, colectomy, ileostomy, date, assessment, action, jp, ex, tube, resection, dilaudid, abdominal, acute, ostomy, comments, flush, fistula, drain, abscess
5	111241	tonic, mandibular, clonic, pedestrian, flolan, struck, mandible, obstipation, frx, fos, fx, seizure, mirapex, pubic, pvi, epilepsy, rami, omfs, car, tendon
6	1.7	bleed, gib, action, assessed, assessment, meq, brbpr, medications, comments, icu, history, ul, pulse, abdominal, balance, bleeding, colonoscopy, response, egd, sbo
7	0.02	infant, caffeine, pn, cares, il, lastname, feeds, cbg, isolette, fio, retractions, wt, baby, mom, spells, murmur, settings, servo, spits, diuril
8	8.1e-4	crrt, hd, cvvh, dialysis, sle, cryptogenic, renal, wound, cvvhd, cirrhosis, lupus, transplant, sirs, cvvhdf, anuric, vac, liver, hemodialysis, paracentesis, crf
9	0.58	cmh, trach, comments, ards, failure, assessed, assessment, ventilation, tube, peg, tracheostomy, meq, ul, respiratory, airway, pna, action, pneumonia, medications, lung
10	0.12	trach, vent, icp, secretions, collar, drain, propofol, coarse, tf, sicu, pt, peg, remains, tragus, eyes, suctioned, goal, psv, intubated, mannitol
11	1.4	evd, hemorrhage, icp, brain, headache, nicardipine, aneurysm, date, mannitol, stroke, mri, cerebral, mass, nimodipine, assessment, ich, vasospasm, checks, sah, cerebellar
12	0.9	transplant, liver, crrt, renal, kidney, failure, cirrhosis, cvvh, micafungin, olt, hepatic, tacrolimus, splenectomy, action, assessment, esld, hepatorenal, arf, dobhoff, acute
13	24.4	neo, iabp, ci, csru, wires, svo, gtt, ntg, lasix, percocet, pacer, endo, milrinone, ct, wean, insulin, drainage, epi, propofol, cabg
14	58.4	stemi, nstemi, rca, lad, ccu, cath, plavix, action, cad, lcx, infarction, myocardial, ami, stent, carotid, cp, coronary, pci, asa, artery
15	1.1	pt, ccu, micu, bipap, sob, wife, confused, cooperative, npn, cath, ew, denies, bed, gu, gtt, nc, lasix, sitter, gi, commode
16	550.6	etoh, valium, ciwa, withdrawal, pancreatitis, abuse, ercp, hiv, alcohol, cholangitis, lipase, action, thiamine, dts, assessment, assessed, tremens, diazepam, delirium, seizures
17	5.8e-3	infant, cares, feeds, mom, active, aquaphor, cc, benign, wt, neonatology, cont, tf, spits, caffeine, ccu, milrinone, pe, voiding, day, remains
18	0.17	trach, vent, secretions, peep, coarse, pt, suctioned, ps, psv, tube, remains, gtt, sputum, settings, abg, lasix, white, tf, mdi, goal
19	244.0	egd, endoscopy, varices, nephrostomy, octreotide, melena, esophageal, bleed, charcot, bleeding, banding, scope, esophagus, cbi, variceal, ent, hematemesis, pheresis, angioedema, protonix
20	13.6	serratia, citrobacter, polymyalgia, infant, duodenal, iabp, enteroscopy, rheumatica, osh, action, jejunal, assessment, iodine, cares, feeds, assessed, vioxx, comments, mom, eclampsia
21	1.3e-6	arrest, pea, versed, fentanyl, vent, peep, intubated, eeg, anoxic, sedation, mcg, sedated, unresponsive, cooling, ett, gtt, secretions, ac, propofol, trach
22	5.2	svg, pod, assessment, bypass, action, cabg, graft, cmh, avr, pedis, coronary, dorsalis, meq, cvicu, aspirin, temporary, medications, artery, response, tibial
23	0.79	pancreatitis, ercp, pancreatic, pancreas, pseudocyst, necrotizing, mrcp, maze, cbd, knee, duct, peri-pancreatic, ligation, cholestasis, biliary, drain, dilaudid, hida, gallstones, necrosis
24	9.7e-3	failure, action, cmh, assessment, assessed, acute, ards, renal, arf, shock, levophed, comments, response, meq, hypotension, ul, pressors, meropenem, pulse, sepsis
25	0.08	sdh, sah, dilantin, seizure, hemorrhage, subdural, assessment, trauma, temporal, action, keppra, comments, cmh, famotidine, fall, epilepticus, subarachnoid, trach, eeg, fracture
26	0.01	lactulose, cirrhosis, liver, encephalopathy, hepatic, ascites, paracentesis, assessed, tips, portal, action, rifaximin, meq, pulse, comments, medications, albumin, hcv, assessment, failure

27	3.9	copd, assessed, history, meq, bipap, action, icu, medications, comments, assessment, pulse, exacerbation, balance, acute, patient, unknown, ed, overdose, nebs, schizophrenia
28	3.0	hd, esrd, dialysis, renal, stage, pd, assessed, chronic, action, kidney, assessment, abscess, failure, crf, meq, bacteremia, pulse, mssa, urosepsis, comments
29	1.0	icd, ep, paced, ablation, vt, ccu, pacer, av, amiodarone, aicd, firing, pacemaker, pacing, lidocaine, block, chb, ppm, dobutamine, milrinone, lido
30	0.09	bmt, gvhd, dka, sct, bactrim, acyclovir, bal, gap, viremia, voriconazole, ards, smx, insulin, tmp, cmv, ketoacidosis, gastroparesis, pcp, vre, viral
31	9.7e-3	aortic, milrinone, valve, endocarditis, nafcillin, mssa, action, bacteremia, mitral, lasix, mvr, assessment, chf, avr, tee, failure, valvuloplasty, ef, meq, septic
32	0.02	thalamic, infant, transplant, action, feeds, liver, assessment, caffeine, cares, esld, landing, cirrhosis, serratia, spells, spits, brainstem, dobhoff, comments, active, mom
33	14.4	infant, cares, feeds, mom, spells, spits, wt, isolette, crib, voiding, active, dev, caffeine, retractions, swaddled, neonatology, pacifier, stooling, fen, pg
34	1.6e-2	pericardial, effusion, cancer, metastatic, lymphoma, chemo, onc, malignant, chemotherapy, tamponade, cell, pleural, oncology, mets, tumor, neoplasm, drain, mass, xrt, follicular
35	0.4	fibrillation, atrial, chf, afib, coumadin, rvr, lasix, action, assessed, heparin, failure, heart, chronic, diastolic, history, assessment, diltiazem, cad, response, acute