THE CURIOSITY CUP 2022
A Global SAS® Student Competition

StayAlert for predicting inpatient Length of Stay (LOS)

Mamiya Adachi, Rehab Meckawy, and Daksh Mukhra,

University of New South Wales (UNSW)

INTRODUCTION

Implementing evidence-based financial principles is integral for the sustainability of the health systems industry. Without proper planning of health spending, systems would collapse with detrimental effects on individual and population health.[1]

The health system payment structure and resource allocation is so challenging given the relatively constrained funding. Health leaders and policy makers would aspire to provide services to all individuals no matter how costly the treatment is. [1] However, decision makers are compelled to wisely allocate resources to achieve the maximum "Value-For-Money". For example, the government is always in a trade-off between treating rare diseases - of expensive management cost - or to diversify the beneficiaries among others with common less costly health conditions. [1] Accordingly, in recent years, health managers and insurance companies have targeted efficiency through *predicting the inpatient Length of Stay (LOS)* as an important parameter to avoid unnecessary health expenditure.

PROBLEM

During the Covid-19 crisis, health systems in both developed and developing countries have witnessed an unprecedented surge in patients' inflow. Emergency departments and hospital aisles were being filled with occupied beds. Moreover, many Covid-19 and non-pandemic cases were hindered from accessing healthcare due to hospitals' overfilled capacity. This strain on hospitals' capacity is greatly attributed to the uncertainty of prospective inpatient Length of Stay (LOS) for each patient.

"The Length of Stay (LOS) represents the interval time between the admission of the patient and his discharge." [2] It is a basic indicator to evaluate the hospitals' performance and quality of healthcare.[2] Given the hospitals' objective to achieve maximum utilization of resources along with timely patient care, hospitals work on increasing patients' flow and abolish the time for non-value added care.[3] Optimal patient flow will not only guarantee effective and efficient care but also minimize the risk of hospital acquired infections and determine the accessibility to inpatient services. [3]

Our project aims to identify LOS predicting factors besides the provision of actionable recommendations to guide health leaders. In case of statistically significant results, our findings will guide efficient use of resources and assist in the preparedness for unexpected surge in demand for hospitals inpatient care. On the other hand, if the investigated LOS connections were not evident, this gives an indication that policy makers should address other LOS predicting factors. To sum up, this project will provide insights on the estimation of LOS at admission time which is necessary for "appropriate planning of care activities".[2]

DATA

The dataset used in this project is a synthetic data for hackathon purpose originally from Analytics Vidhya, publicly shared on AV : Healthcare Analytics II | Kaggle. The dataset consists of the length of stay (LOS) category as the outcome variable, and other variables are features (17 variables) related to patient (e.g. patient ID, age, type of admission) and hospital (e.g. hospital code, region, bed grade). The records (rows) in the dataset are held per admission case, labeled by case ID. The original dataset is provided separately as default into training (318,438 cases) and test (137,057 cases) sets for both model development and performance assessment purposes.

DATA CLEANING/VALIDATION

Before any models can be trained, we must clean and impute the data, conduct exploratory analysis, and do feature engineering. Data was first checked for erroneous and missing values as well as mismatching data types. Only two variables (Bed grade and City patient code) contained missing values, which comprised approximately 1% of responses for each variable. As such missing values were removed which has a negligible effect on the distribution for the respective variables. The data provided contained multiple categorical features with a high number of subclasses (25+) such as city codes for hospitals and patients. Apart from length of stay such variables were nominal and hence were still encoded using traditional dummy variables. As this resulted in high dimensionality we used variable selection in our modeling. Once data was cleaned, we conducted exploratory data analysis.

EXPLORATORY DATA ANALYSIS

Figure 1

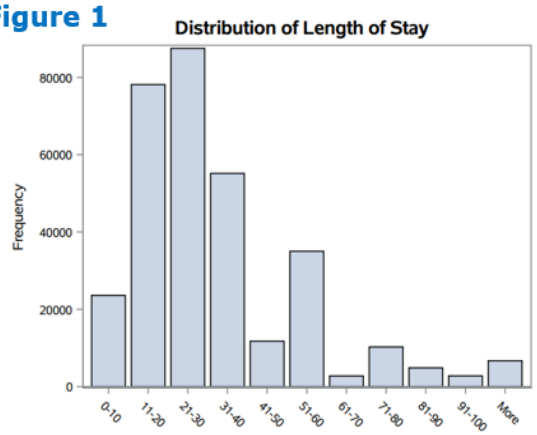


Figure 2

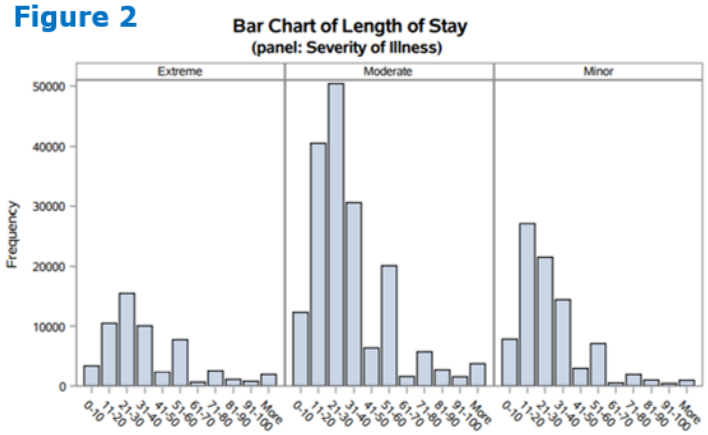


Figure 3

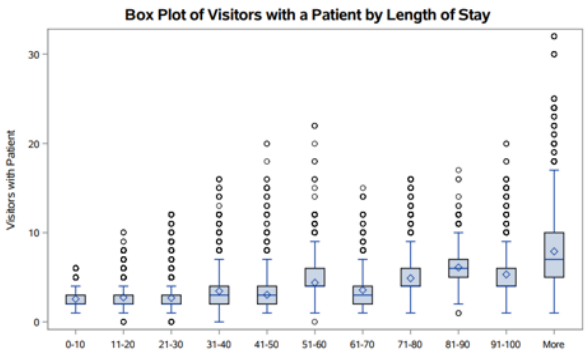
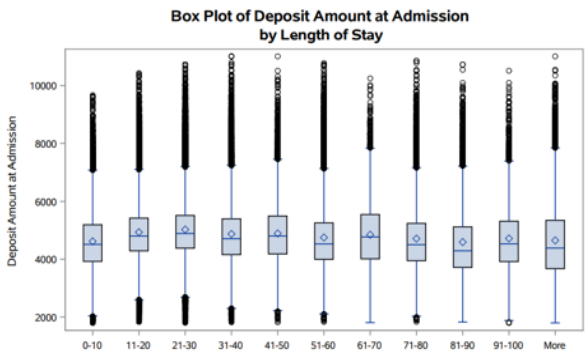


Figure 4



First analyzing the distribution of length of stay (Figure 1), we can see this distribution is highly right skewed and this may be problematic for classification. Where in this case the minority classes "patients with higher length of stay" would want to be predicted with high accuracy so hospitals can optimize resources. Looking at figure 3 from the boxplots we can see that the median number of visitors for patients with length of stay greater than 81 is higher than those below, suggesting a weak correlation between the two factors. Figures 2 and 4 respectively showcase the distribution of severity and the distribution of deposit amount for each class in length of stay; however there are no significant shifts seen from the visualizations. Furthermore it was found apart from Visitors with patients there is no combination of variables that showcased a significant effect on length of stay. This may seem counterintuitive but as this data set was not a real-world data set unlikely results are bound to occur. However our statistical models may be able to uncover some hidden relationships.

ANALYSIS

The primary goal in modeling was to investigate the relationship between LOS (length of stay), a multiclass categorical variable and various factors such as admission deposit and number of visitors. Predictive analysis was done using Multinomial logistic regression and random forest. And both models used a 70-30 training and test split.

Multinomial Logistic Regression:

As the data in LOS may be regarded as ordinal, one may be inclined to perform ordinal regression. However the proportional odds assumption must be satisfied in order to proceed with it. This assumption was tested in SAS Studio indicating the data did not satisfy said assumption hence multinomial regression was used. The regression model was coded using **proc hpgenselect** in SAS Studio which automatically converted categorical factors into dummy variables and standardized continuous variables. LASSO (Least Absolute Shrinkage and Selection Operator) was used to perform dimensionality reduction.

The frequency of correctly classified observations was 27.59 %, indicating the model has not performed well.

Only admission deposit remained as a predictor for all regression equations in the multinomial regression. The results show the logistic coefficient for each predictor variable in our case admission deposit for each alternative category of the outcome variable (LOS); not including the reference category ("LOS 0-10"). As such below is an interpretation of the factors with highest magnitude coefficients. All full results in Figure 6 in appendix.

- A one-unit increase in the variable admission deposit is associated with a 0.144×10^{-3} decrease in the relative log odds of having LOS 81-90 vs. LOS 1-10.
- A one-unit increase in the variable admission deposit is associated with a 0.10310×10^{-3} increase in the relative log odds of having LOS 21-30 vs. LOS 1-10.

As one can see the magnitude of the coefficients is extremely small indicating only a weak relationship is present between admission deposit and LOS, and the other variables seemingly have no relationship at all.

Random forest:

The Random Forest is an ensemble method which takes multiple individual learning models known as weak learners and combines them to produce an aggregate model that is more powerful than any of its individual learning models alone. The random forest model was coded in SAS Studio using **proc hpforest**. The frequency of correctly classified observations is 38.79% which is significantly higher than the regression. This is due to the inherent tree structure that can deal with high dimensional data sets. The most significant feature in

deciding LOS (see appendix Figure 7 for further information) from the model was the number of visitors which matched conclusions reached from our exploratory data analysis unlike the regression.

SUGGESTIONS FOR FUTURE STUDIES

Our main recommendation for future studies is to explore country-related associations between Length of Stay (LOS) and the two variables analyzed above; visitor number and admission deposit.

Country-recognizable datasets would take into consideration contextual factors in terms of health administration and patients' behaviors when it comes to interaction with health service delivery. For example, in LMICs where most patients rely on subsidized governmental sectors, visitor number is neither restricted nor observed. This is usually associated with hospital wards' overcrowding (risk of infection) and entry of unhealthy diets to the admitted patients. The latter behaviors would lead to prolongation of LOS since patients' recovery and treatment process would be slowed. On the contrary, in HICs, these visitors' behaviors are absent and likely to avoid factors that would have led to preventable prolongation of LOS.

Additionally, country-related associations between LOS and admission deposit would take into account the health system's payment modality. Admission deposits embody a prepayment policy where the latter would indirectly incentivize health administrators to prolong LOS since more profit is adding up to the discharge fee. However, adopting post-care payment modality carries an uncertainty element for hospital administrators that the patient is capable of affording care and this would encourage providers to shorten the LOS. In other words, admission deposit gives security to hospital managers that admitted patients are affording care and prolonging LOS would incur further revenues. Therefore, country-related associations would give a clue of countries' payment structure and how the latter incentivize or discentivize care providers to unnecessarily shorten or prolong LOS.

CONCLUSION

Our analysis revealed that visitor number and admission deposits are the most worthy of consideration when predicting LOS. However, more real world data would need to be collected particularly for exploring country-related associations that take into account health system payment structure and public attitude into account.

REFERENCES

Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. (2016). Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association, 23*(e1), e2-e10.

Meckawy, R. (2021). *Demonstrated Mastery of MIHM Competencies*. Arizona State University. MIHM Portofolio

Rachda Naila, M., Caulier, P., Chaabane, S., Chraibi, A., & Piechowiak, S. (2020). *Using machine learning models to predict the length of stay in a hospital setting*.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Mamiya Adachi - m.adachi@student.unsw.edu.au
Rehab Meckawy - rehabmeckawy@gmail.com
Daksh Mukhra - dakshmukhra1@gmail.com

APPENDIX

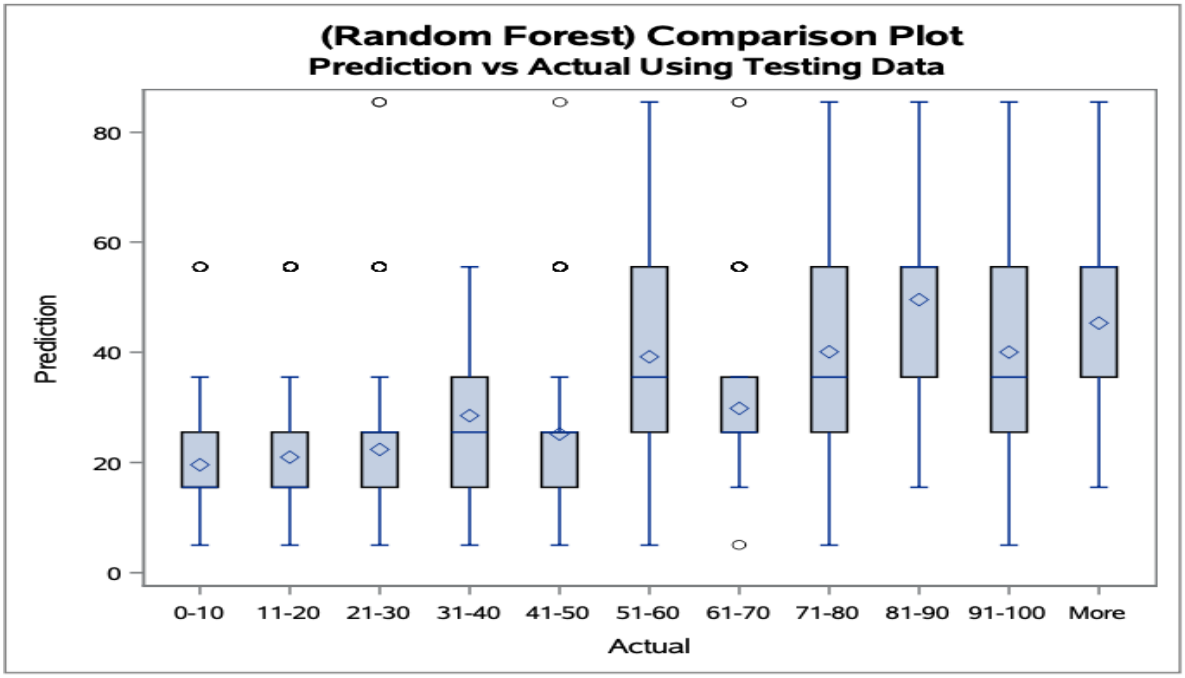


Figure 5

Multinomial Logistic Regression Coefficients

Parameter Estimates			
Parameter	**Stay_sort**	**DF**	**Estimate**
Intercept	More	1	-1.273861
Intercept	91-100	1	-2.164367
Intercept	81-90	1	-1.607352
Intercept	71-80	1	-0.848691
Intercept	61-70	1	-1.967008
Intercept	51-60	1	0.377950
Intercept	41-50	1	-0.691062
Intercept	31-40	1	0.873205
Intercept	21-30	1	1.334342
Intercept	11-20	1	1.214150
Admission_Deposit	More	1	0
Admission_Deposit	91-100	1	0
Admission_Deposit	81-90	1	0
Admission_Deposit	71-80	1	0
Admission_Deposit	61-70	1	0
Admission_Deposit	51-60	1	-0.000000193
Admission_Deposit	41-50	1	0
Admission_Deposit	31-40	1	0.000001487
Admission_Deposit	21-30	1	0.000006751
Admission_Deposit	11-20	1	0.000001292

Figure 6

Random Forest Model

Friday, January 28,

Loss Reduction Variable Importance					
Variable	Number of Rules	Gini	OOB Gini	Margin	OOB Margin
Visitors with Patient	8973	0.048998	0.01920	0.018128	-0.00446
Ward_Type_sort	1677	0.011016	0.00484	0.005659	0.00117
Typ_Adm_sort	1169	0.005662	0.00115	0.003007	-0.00015
Bed_Grade_sort	1902	0.008560	0.00082	0.005289	-0.00032
Svr_Illn_sort	1617	0.005390	-0.00101	0.001005	-0.00351
Department_sort	977	0.002175	-0.00161	0.001139	-0.00159
Ward_FclCd_sort	1452	0.004369	-0.00185	0.001919	-0.00256
HospReg_sort	1201	0.002730	-0.00205	0.001381	-0.00196
HospTyp_sort	1488	0.003419	-0.00264	0.002092	-0.00233
City_Code_Hospital	1447	0.003953	-0.00276	0.002043	-0.00301
CityCd_Pt_sort	1474	0.003544	-0.00434	0.002469	-0.00389
Hospital_code	1786	0.005052	-0.00452	0.003272	-0.00404
Age_sort	2139	0.004684	-0.00662	0.002965	-0.00521
Available Extra Rooms in Hospita	7419	0.014174	-0.01239	0.012143	-0.00771
Admission_Deposit	10758	0.025353	-0.02206	0.022459	-0.01465

Figure 7