**HAP 880 Final Project Report in**

**Predicting Hospital Length of Stay on MIMIC III Data**

Merin Joy – G01118158

Volgenau School of Engineering, George Mason University, 4400 University Dr,

Fairfax, VA, 22030

**Abstract**

This study deals with predicting the hospital length of stay using Logistic Regression, Naïve Bayes classifier and Random Forest model with various diagnoses based on selected characteristics. The models are trained to predict whether the patient’s length of stay will be long(>median) or short(≤median). Here, longer and shorter stay is determined using the median of length of stay evaluated using the data in MIMIC III database. Required tables are drawn from the MIMIC III database and all codes are written in Python.

**Key Words:** MIMIC (Medical Information Mart for Intensive Care), Logistic Regression, Naïve Bayes, Random Forest, Python.

**1. Introduction**

Predicting the hospital length of stay is of significant value for production, planning and management within a hospital. If we have the resources to estimate how much longer a person stays in a hospital, it can help us in assisting in the scheduling the usage of wards and availability of hospital beds including the scheduling of elective surgeries based on upcoming vacancies of hospital beds. Also, cost related consumption can be explained using the length of stay. About 80 to 90 percent of the hospital costs between patients corresponds to length of stay.

An estimate in length of stay will assist hospital administrators, clinicians, patients and payers. In the case of hospitals, its recommended to have an optimized use of beds for better care, for clinicians predictive models can provide adjunctive clinical decision support. For patients, improved planning, care and medications improves the quality of care and support provided. Also, payers are continuously seeking tools to increase cost analysis and prediction. Also, a value based care approach is growing increasingly in the field of hospital and health care.

In predicting the length of stay, one of the many challenges one can come across is the numerous factors responsible for length of stay. The complexity of the problem causes us to think about this problem more. The challenge stays in determining the multiple factors and their complex inter-relationships in order to estimate length of stay.

In this project, MIMIC III database was utilized for predictive modeling. In database, although de-identified, contains detailed information regarding the clinical care of patients. Hence its is required to treat with appropriate care and respect. Prior to requesting access to MIMIC, an online course called “Data or Specimens Only Research” is required to be completed. MIMIC contains the detailed data of patients admitted in the Beth Israel Deaconess Medical Center in Boston and was developed by the MIT Lab for Computational Physiology.

**2. Datasets**

The huge part of this project was to better analyze the tables present in MIMIC database, especially the Electronic Medical Records(EMRs). MIMIC contains more than 50000 thousand records of patients admitted in ICUs. The raw data available on MIMIC is in table format with metadata dictionaries and an online query tool.

The MIMIC database covers several types of information:

* Admissions: details about a patients admission events
* Discharges: details of patients leaving hospital
* Caregivers: details of what type of staff cared for a patient during their stay
* Chart information: details of a patient’s medical chart
* Labevents: details of lab tests and events
* Diagnoses: details of the various types of diagnoses performed on a person
* Procedures: details of the procedures carried out on a person.
* Patients: details of a person’s demographic information, notes and records

**3. Models**

1. Logistic Regression

Logistic Regression is the most appropriate model to conduct when the dependent variable is binary. Logistic regression is used to describe data and explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Logistic regression can be widely used in various fields, including machine learning, most medical fields, and social sciences. An example will be the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression. Many more medical scales used to assess severity of a patient have been developed using logistic regression. Logistic regression may be used to predict the risk of developing a given disease based on observed characteristics of the patient.

1. Naïve Bayes

In [machine learning](https://en.wikipedia.org/wiki/Machine_learning), naive Bayes classifiers are a family of simple "[probabilistic classifiers](https://en.wikipedia.org/wiki/Probabilistic_classifier)" based on applying [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) with strong (naive) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features. Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem. [Maximum-likelihood](https://en.wikipedia.org/wiki/Maximum-likelihood_estimation) training can be done by evaluating a [closed-form expression](https://en.wikipedia.org/wiki/Closed-form_expression), which takes [linear time](https://en.wikipedia.org/wiki/Linear_time) than by expensive [iterative approximation](https://en.wikipedia.org/wiki/Iterative_method) as used for many other types of classifiers. With appropriate pre-processing, it is competitive in complex domain specific categories with more advanced methods including [support vector machines](https://en.wikipedia.org/wiki/Support_vector_machine). It also finds application in automatic [medical diagnosis](https://en.wikipedia.org/wiki/Medical_diagnosis).

1. Random Forest

Random Forest is an ensemble learning model that is capable of performing both regression and classification tasks. Ensemble learning model refers to n number of weaker learning models together combined to form a powerful one. The standard weaker learner like a decision tree is used for a random forest. An input in a decision tree is entered at the top and as it traverses down the tree the data gets put into smaller and smaller sets or classes. In the case of classification, a mean prediction by averaging the results of individual trees is generated.

A decision tree is comprised of nodes and splits of the data. ​The tree starts with all training data residing in the first node.​ An initial split is made using a predictor variable, segmenting the data into 2 or more child nodes. Splits can then be made from the child nodes. A terminal node is one where no more splits are made. Predictions are made based on the make-up of terminal nodes.​

**4. Experimentations**

1. Median LOS

The objective of this project is to estimate whether a person will have a “longer” stay (>median) or a “short” stay (≤median) using various modeling algorithms. The median LOS value obtained here is around 7.8 days. Median was chosen because it is the middle numerical value separating all of the individuals into two halves when ranking (high to low) each admission to a Center by their LOS value. One half of all the residents have a LOS value greater than the median and the others of the individuals have a LOS value less than the median. The median is used rather than the mean (or average) because the LOS average can be misleading by having a few residents with very long LOS.

For example, assuming a hospital has 5 admissions, each with LOS of 5, 10, 15, 20 and 100 days respectively. The median will be 15 days (the value where half resident have lower and half have higher values), while the average (or mean) will be 30 days [i.e., (5+10+15+20+100)/5]. The mean (or average) of 30 days is higher than the LOS for 4 of the 5 residents while the median is the exact mid-point.

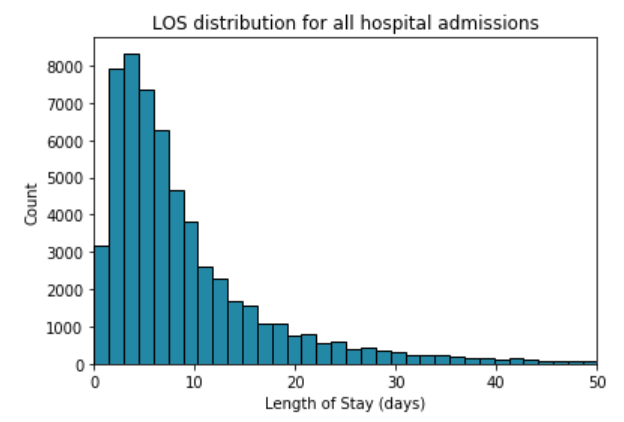


Fig 4.1

1. Datasets

Seven tables from the MIMIC III database were combined and used to create the input attributes:

**Admissions:** hadm id, subject id, admittime, dischtime, ethnicity, admission type, insurance, marital status

**CPT Events:** costcenter, cpt cd

**ICU Stays:** ﬁrst careunit, last careunit

**Services:** curr service

**Patients:** gender, dob

**Procedures ICD:** seq num

**Diagnoses ICD:** icd9 code (truncated to ﬁrst 3 characters)

For the dependent variable Length of Stay (LOS), the value was calculated from discharge time and admission time of the admissions table.

1. Preprocessing

As expected MIMIC III database tables were pretty raw and required data pre-processing to obtain a much cleaner and model likable format. All seven of the above mentioned tables were pre-processed into simpler and understandable format.

1. Admissions

Firstly, the dependent variable LOS was calculated using the admit time and discharge time. LOS value below zero were removed since it did not make sense to have admit time greater than discharge time. Ethnicity column contains 41 categories which were converted to six major categories. Null values in Marital Status were converted to unknown or default category. The categorical variables were later on changed to columns having 0s and 1s using pandas get dummies method.

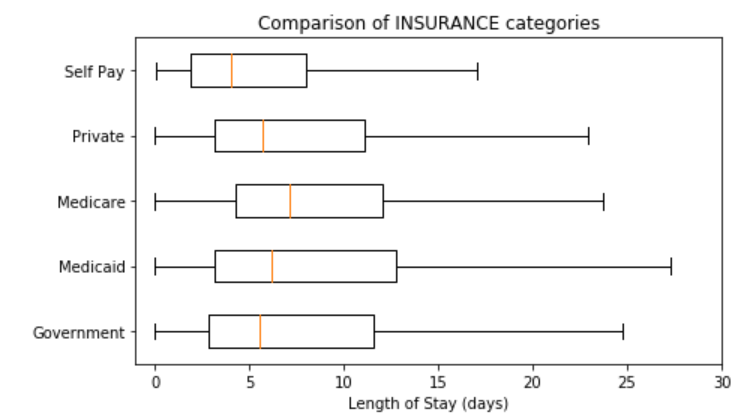
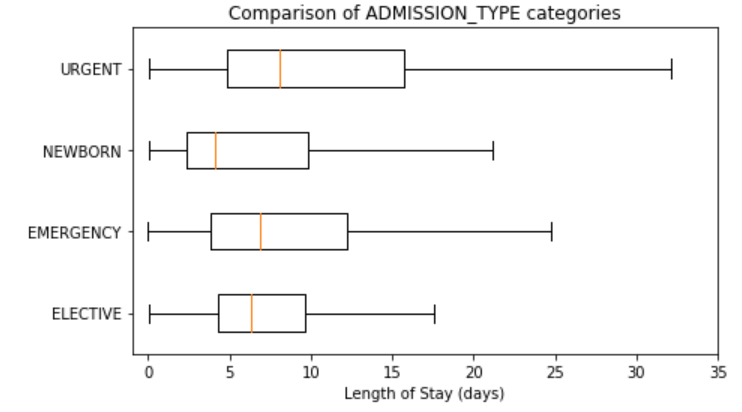


Fig 4.2

1. CPT Events

Contains current procedural terminology (CPT) codes, which facilitate billing for procedures performed on patients. Costcenter is the cost center which billed for the corresponding CPT codes. There are two possible cost centers: ‘ICU’ and ‘Resp’. ‘Resp’ codes correspond to mechanical or non-invasive ventilation and were billed by the respiratory therapist. ‘ICU’ codes correspond to the procedures billed for by the ICU. CPT Events table provides some information on the nature of the patient’s condition, having just two possible values ICU or RESP. A CPT sequence number was created to establish the number of CPT codes associated to an admission. Categories were then changed to columns using dummies.

1. ICU Stays

The ﬁrst careunit and last careunit inputs act to provide a proxy for information about the broad type of health conditions and categories of care a patient requires. These careunits can have a bearing to LOS. They were converted to columns with dummy values.

1. Services

Although, previous service cannot help us much since most of the values in the column is not recorded. Current service could be of use as a determining factor for length of stay estimation. These services were converted to columns using dummy values.

1. Patients

From the patients table, demographic values like gender and DOB of a person are extracted. Gender can converted to columns using pandas get dummies and DOB is used for determining the age of the person.

1. Procedures ICD

Procedure number was generated using the maximum of sequence number from Procedures ICD table. This number of procedures will give us an insight on how many procedures were applied on a patient during the admission event.

1. Diagnoses ICD

ICD9 codes were separated into 16 different super categories and coded to form columns containing dummies values.

All the above mentioned tables were then merged together to obtain the final desired table of input attributes. Attribute age in MIMIC III dataset can be calculated using difference of admit time and date of birth, but patients who are older than 89 years old at any time in the database have had their date of birth shifted to obscure their age and comply with HIPAA. The patient’s age at their first admission was determined and the date of birth was then set to exactly 300 years before their first admission. In this case, correct date of birth can be obtained by using the first admit time of each patient and subtracting it by corresponding DOB of patient. An assumption was made for patients who have age less than zero, their age was converted to 90.

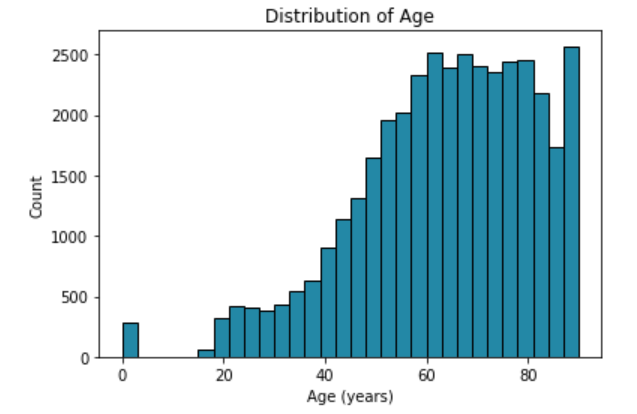


Fig 4.3

**5.Evaluation Metrics**

Evaluation Metrics are the metrics used for performance evaluation of a model. Below mentioned set of metrics with accuracy that are mainly considered in determining which model performs better from the rest of them. Confusion Matrix is also used as a metric in determining the model efficiency.

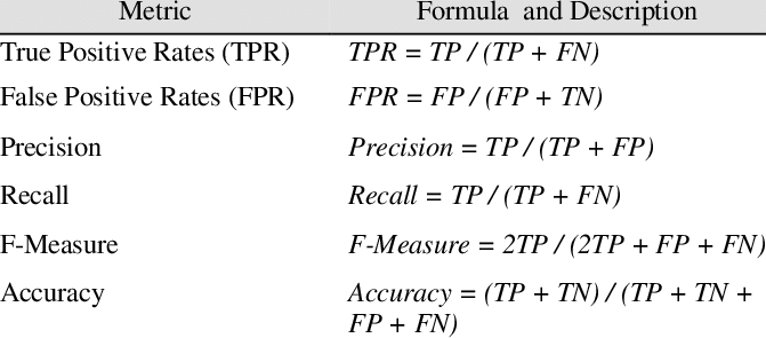


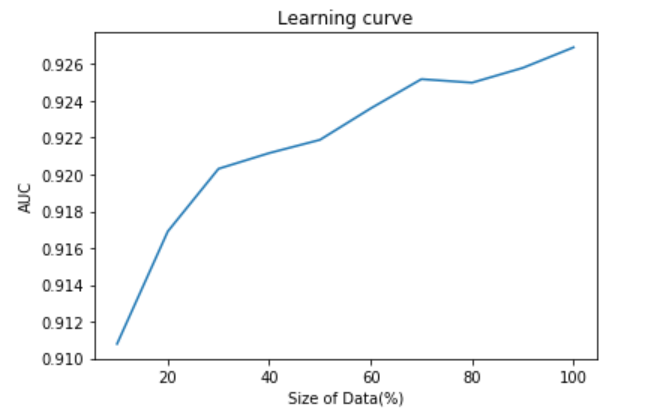
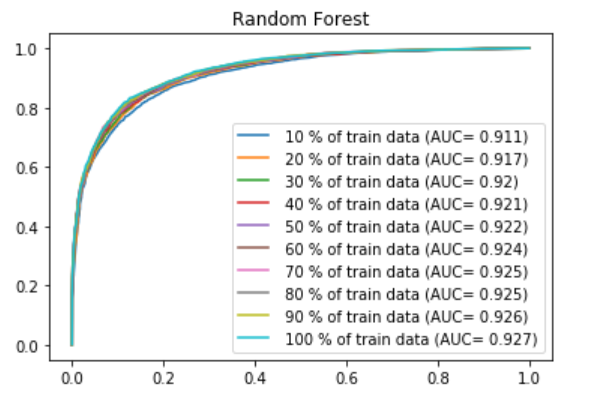
Fig 5.1

**6. Results**

1. Random Forest

The random forest algorithm was run using different methods. With the number of estimators as 400, accuracy was 83.93%. Feature selection was done using chi square and mutual information values. Almost similar accuracies of 83.70% and 84.48% were obtained for mutual information and chi square respectively. Also, 84.75% accuracy was achieved on 90% of training data.

Best Results – 84.75% accuracy



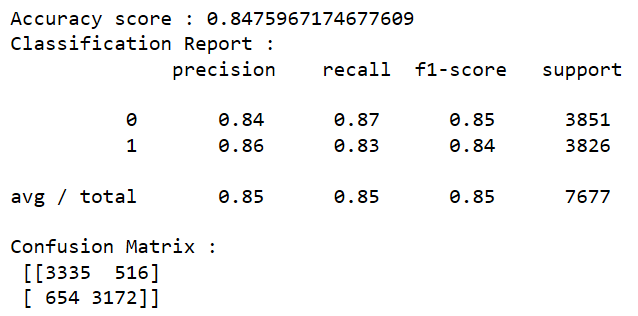
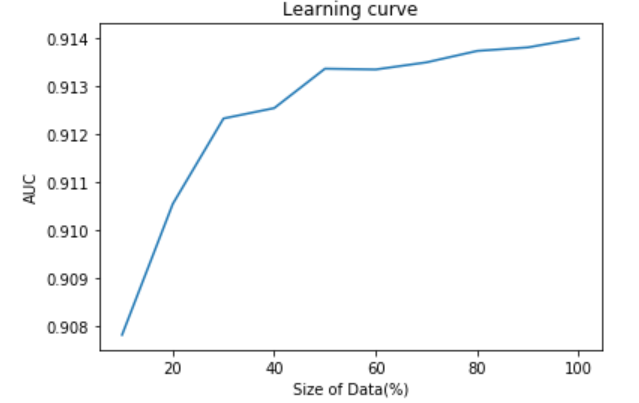
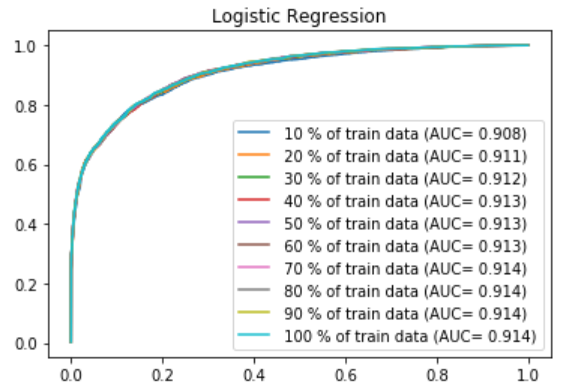


Fig 6.1

1. Logistic Regression

Default logistic model gave an accuracy of 82.16%. Feature selection using chi square and mutual information gave 82.18% and 82.94% respectively. Accuracy of 82.98% was obtained 80% of the training data.

Best Results – 82.98% accuracy



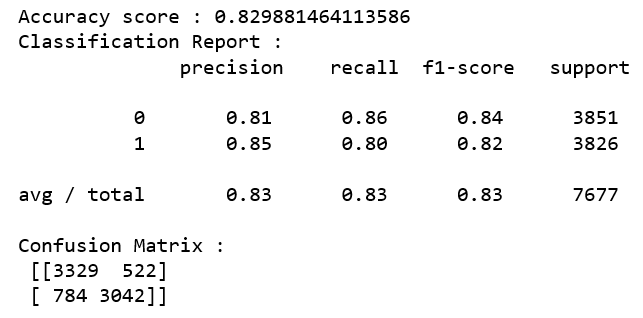
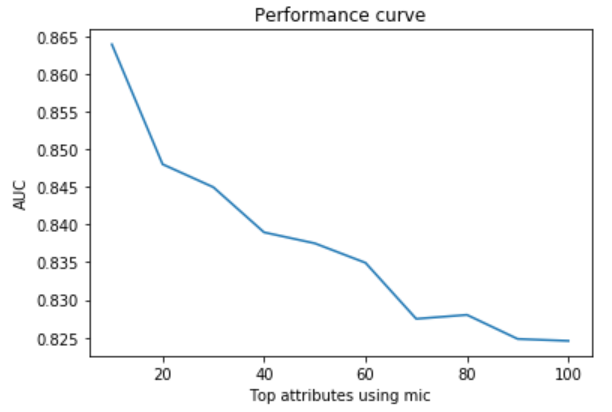
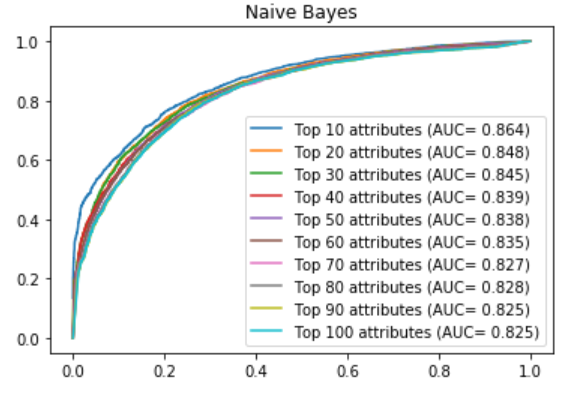


Fig 6.2

1. Naïve Bayes

Default naïve bayes model gave a very low accuracy of 68.64% but using feature selection of mutual information on top 10 attributes gives an accuracy of 77.06%. Also, chi square value on top 10 attributes give 76.33% accuracy. Using only ten percent of the training dataset gives an accuracy of 73.97%.

Best Results – 77.06% accuracy



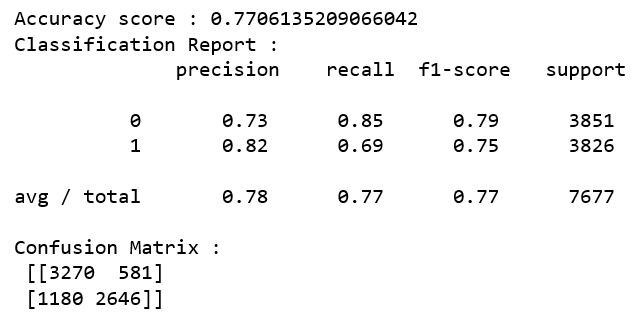


Fig 6.3

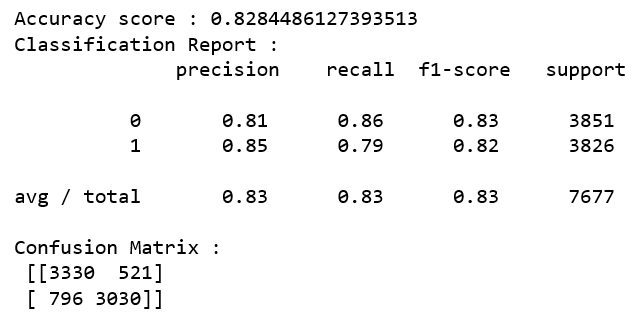
**6.1 Grid search and Cross validation Results**

In order to have better accuracy and results, two types of grid search techniques are used: Random search and Grid Search with 5-fold cross validation of training dataset. The results obtained from the random search are used for grid search, thus increasing the efficiency of the model. Later, the results from default model and best search models are compared using accuracies in terms of efficiency.

1. Logistic Regression

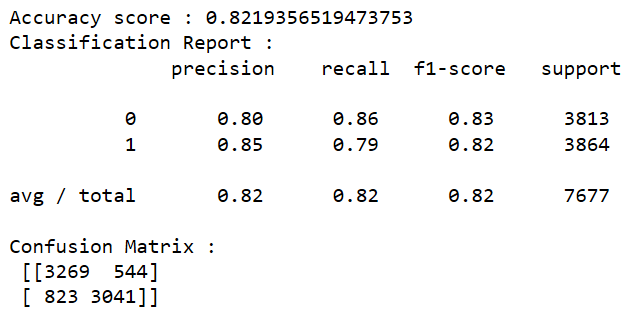
**RandomSearchCV**





**GridSearchCV**

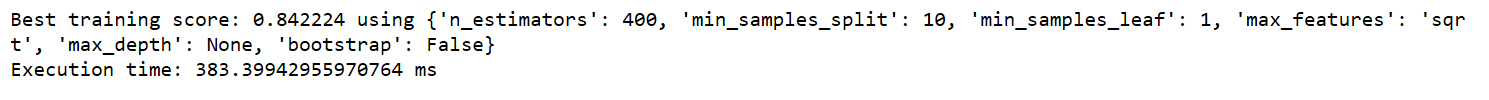


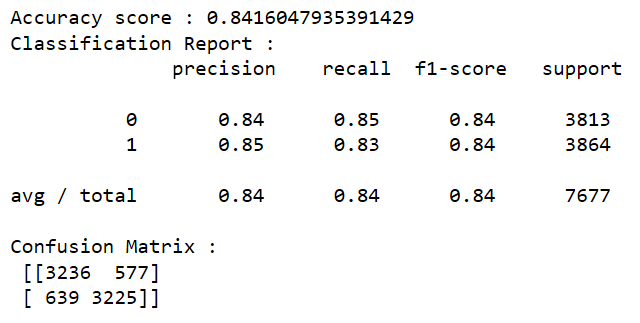


We can see that from the default model, random search cross validation has an accuracy improvement of 0.68% and for grid search its 0.03%.

1. Random Forest

**RandomSearchCV**





Around 2.51% improvement from the default model.

**7. Future Work**

1. Including all of the ICD9 codes. All have a potential to have a bearing on Length of Stay.
2. Inclusion of the ﬁrst procedure ICD9 code, that is the ICD9 code of the ﬁrst procedure carried out during an admission, may also prove beneﬁcial.
3. Taking into consideration ICU length of stay. This would have additional high clinical value as the cost per night for ICU-based care is typically multiples of the cost per night of a non-ICU ward bed, and so estimating ICU LOS has a per night greater implication for resource planning.
4. The inclusion of speciﬁc physiological data measurements, care-giver observations, lab and imaging test results and medication information.

**References**

[1].Koehrsen, Will (January 10, 2018).Hyperparameter Tuning the Random Forest in Python. Retrieved from <https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>

[2]. Nevrekar, Akshay (November 1, 2018) Random Forest using GridSearchCV.

Retrieved from <https://www.kaggle.com/sociopath00/random-forest-using-gridsearchcv>

[3]. Maria Kelly, Linda Sharp, Fiona Dwane, Tracy Kelleher, and Harry Comber. Factors predicting hospital length-of-stay and readmission after colorectal resection: a population-based study of elective and emergency admissions. BMC Health Services Research, 12(1):77, 2012.