STAT3612 Statistical Machine Learning

30-day All-Cause Hospital Readmission Prediction

Final report (GROUP 11)

Name: CHAN KWOK CHEUNG 3035821556, CHIU HOI KIT MARCO 3035781976, CHAN YIN KEI 3035787499, LI CHI TAT 3035783455, FAN ANGUS CHUNG MAN 3035995539

Abstract

Statistical Machine Learning (SML) has gained popularity in the Big Data era due to its ability to handle large and complex datasets. This project aims to look at the efficacy of several SML techniques in predicting hospital readmissions and to select the best model with the highest accuracy. The study has advanced gradually, effectively forecasting hospital readmission rates using Gradient Boosting and Random Forest models. Both models and their performance have been evaluated. This paper contains extensive experimental data as well as an update on the current state of the project. In the meantime, the development and results have been positive, revealing that these SML models can predict hospital readmissions with high accuracy. However, there is still room for improvement in order to improve the models' performance. In the future, the models will be updated and enhanced to improve their accuracy and usefulness in anticipating readmission rates. This study aims to reduce preventable hospital readmissions and improve patient care management by leveraging the capabilities of SML techniques and constantly refining the models. The findings of the study have the potential to affect healthcare decision-making and improve patient outcomes.

1. Introduction

1.1 Background

Electronic health records (EHRs) have enormous promise for improving healthcare across a wide range of fields, including resource allocation, patient prioritizing, and hospital demand forecasts. The goal of this research project is to investigate the construction of statistical machine learning models that use data inputs to maximize prediction accuracy for risk assessment tasks. We will investigate the effectiveness of models in making robust and dependable predictions. We hope to determine the best way for reliable risk prediction in healthcare settings by studying and comparing the performance of these strategies. This attempt has the potential to dramatically advance the field by improving decision-making processes and, eventually, improving patient outcomes.

1.2 Aims

The project intends to solve the issue of needless hospital readmissions and improve patient care while increasing overall healthcare system efficiency. Our goal is to create models that can help improve patient participation and self-management, as well as promote medical research and policy development. To do this, we will create and deploy two predictive models to identify significant characteristics and predictors for the aforementioned use cases.

We will use a variety of data inputs to create these models, including patient demographics, medical history, vital signs, test results, and clinical notes. We can determine the most relevant elements that lead to patient death and length of stay by studying and comprehending the patterns and relationships within this data. These variables can then be employed as predictors in our models, allowing us to make accurate and dependable predictions.

These predictive models have the potential to transform healthcare decision-making processes. We can give healthcare professionals useful insights and tools to improve patient care, optimize resource allocation, and eventually improve the overall efficiency of the healthcare system by employing modern statistical and machine-learning approaches.

1.3 Datasets

This project's training dataset is the Electronic Health Record Data (MIMIC-IV v1.0).

This dataset contains health information from over 250,000 individuals admitted to Boston's Beth Israel Deaconess Medical Center between 2008 and 2019. Demographics, vital signs, laboratory results, prescriptions, procedures, diagnoses, and clinical comments are all included in the dataset.

MIMIC-IV is known for its emphasis on intensive care unit (ICU) patients. The dataset contains detailed data from the ICU setting, allowing researchers to investigate critical care scenarios and interventions.

This dataset, we believe, can aid in the development of statistical machine-learning algorithms for reducing needless hospital readmissions.

1.4 Planning

We appreciate the importance of selecting models with high prediction accuracy in our quest to prevent needless hospital readmissions. To do this, we examined linear and non-linear models and discovered that non-linear models have superior prediction accuracy. As a result, we chose two non-linear models for our research: Gradient Boosting and Random Forest.

Gradient Boosting is a machine learning approach that combines several weak predictive models, such as decision trees, to produce a more powerful and accurate prediction model. It works by training these weak models consecutively, with each successive model focusing on rectifying the errors generated by the prior models. The end consequence of this repeated procedure is a very accurate and robust predictive model.

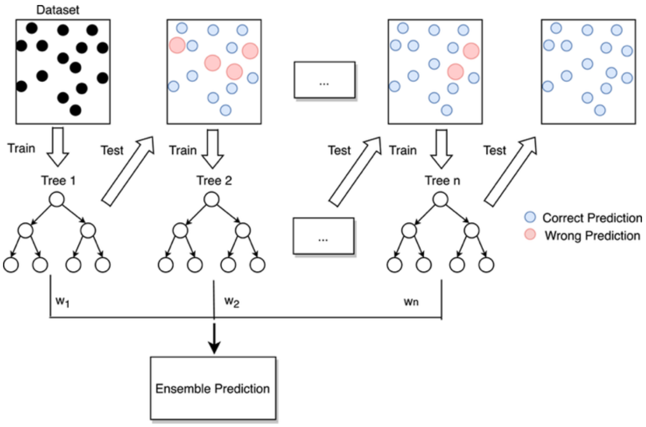
Random Forest, on the other hand, is an ensemble learning method that builds several decision trees and then combines their forecasts to produce correct predictions. Each decision tree is constructed using a randomly chosen subset of the data and features, ensuring diversity and lowering the danger of overfitting. Random Forest gives a dependable and accurate forecast by aggregating the predictions of several decision trees.

We chose these two non-linear models to capitalize on their capabilities in capturing complex relationships and patterns in data. This will allow us to make more accurate predictions for risk assessment tasks like identifying patients who are at a higher risk of hospital readmission. Finally, we want to create models that can considerably help to reduce needless hospital readmissions and improve patient care outcomes.

2. Methods and Result

2.1 Gradient Boosting

Gradient Boosting is an ensemble machine learning technique that combines weak predictive models, such as decision trees, to create a strong predictive model. It sequentially adds models to the ensemble, with each new model correcting the mistakes made by the previous ones. The models are trained to predict the residuals of the ensemble, reducing errors through gradient descent optimization. Gradient Boosting is versatile, handling various data types and missing values effectively. It provides accurate predictions by capturing complex relationships in the data. However, it requires proper regularization to avoid overfitting and can be computationally expensive.



(Zhang, Tao & Lin, Wuyin & Vogelmann, Andrew & Zhang, Minghua & Xie, Shaocheng & Qin, Yi & Golaz, Jean-Christophe. ,2021)

2.1.1 Data Preprocessing

The initial phase of the data preprocessing procedure involves loading the raw preprocessed data. This data, which resides in a pickle file, namely ehr_preprocessed_seq_by_day_cat_embedding. pkl file, once loaded into the data variable, reveals a Python dictionary, feat_dict. This dictionary consists of entries, each corresponding to a unique patient. Each entry comprises a unique patient ID serving as the key, and the associated value is an array of Electronic Health Records features pertinent to the respective patient.

Subsequently, the procedure moves onto the loading of CSV data. Three CSV files are imported into the system, named train.csv, valid.csv, and test.csv. These files encapsulate more comprehensive information about the patients, which may not be encompassed within the EHR features.

Following is the deduplication process comes into effect, where potential duplicate rows within the CSV datasets are eliminated. The determination of duplication is based on the first column, which is assumed to be a unique key, presumably the patient ID. The drop_duplicates function is utilized for this purpose. The cleaned, deduplicated data is subsequently stored back to new CSV files. This integral step ensures that the model is not biased by duplicate entries and that each patient's data is represented singularly.

Post deduplication, the procedure advances towards filtering the PKL data. This phase involves the filtration of the feat_dict dictionary from the pre-processed PKL file to solely include records possessing keys that are also present in the deduplicated CSV files. This phase ensures that the PKL and CSV data are aligned, thereby only containing information about the same set of patients. The filtered data is then stored as new PKL files.

The final preprocessing phase appends an additional feature to each record within the filtered PKL data. This feature, extracted from the 12th column of the CSV data, is appended only when the key in the CSV data matches the key in the PKL data. This step ensures the correct feature is appended to the corresponding patient's record. The resulting enhanced datasets are stored as new PKL files.

This marks the conclusion of the data preprocessing phase. The output of this procedure is a series of PKL files, each encapsulating the data pertaining to a set of patients. Each record now comprises the original EHR features, augmented with an additional feature from the CSV file. The data is now in a suitable format and is ready to be fed into a machine learning model for training.

2.1.2 Approach

In this project, gradient boosting model was utilized for the prediction of 30-day hospital readmissions. The model was trained using data from the updated PKL files by using XGBoost, which could enhance the speed and performance regarding the large dataset. To predict the readmission probability of patients, the model assigned a predicted probability to each row of records for each individual patient. Consequently, three distinct CSV files were generated. These files represented the mean, maximum, and final day's probability for each patient and served as the results for our study.

Since the gradient boosting model itself is well-equipped to handle high dimensional data, so we did not do such a process to reduce the dimensionality of the data. Similarly, the Synthetic Minority Over-sampling Technique (SMOTE) was initially considered to handle the class imbalance issue within the data. Still, it seemed gradient boosting model were found to adequately handle this issue. No significant improvement was observed, and it was not incorporated into the model eventually.

Fine-tuning was applied, while some hyperparameters such as reg_alpha and reg_lambda were included to enhance model's performance and mitigate potential overfitting. Additionally, feature selection was carried out with a goal to identify important features to build a reduced model, hence further prevent overfitting. Random state was set to 32 throughout the process to ensure consistent performance measure throughout the process.

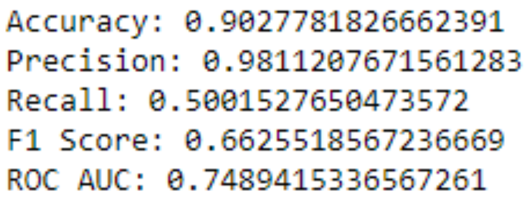
A comparative analysis between the model with and without the application of fine-tuning was conducted. The results revealed a noteworthy difference in performance, specifically the accuracy of the model improved by an approximate 10%. Consequently, it can be concluded that the implementation of fine-tuning strategies can significantly augment the accuracy of the model.

2.1.3 Result

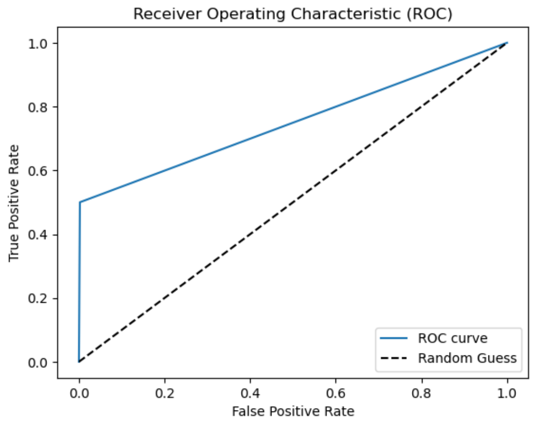
The results obtained from the analysis provide significant insights into the model's performance. The model exhibits an exceptional capacity for predicting the probability of patient readmission within a 30-day period in the valid set by executing the binary classification.

The model achieved 0.903 accuracy in the valid set, and even reached an outstanding 0.981 for the precision score, proving a strong ability to make correct positive predictions overall, or in this case, patients who are most likely to be readmitted. Considering the primary focus is on patient health, a model with higher precision would be particularly beneficial for our goal, thereby enhancing the efficiency of health interventions.

While the recall and F1 scores indicated average to moderate performance, the ROC AUC score of 0.749 shows a certain strength of discriminatory power in distinguishing between readmission cases and non-readmission cases.



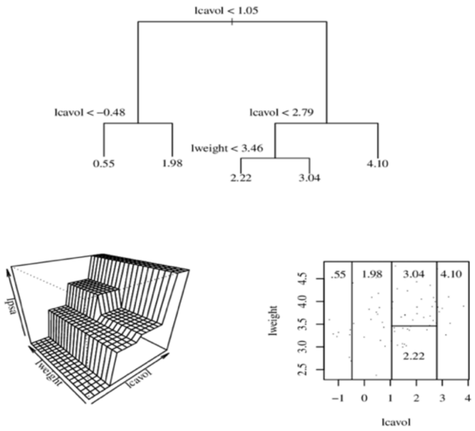
```
Accuracy: 0.9027781826662391
Precision: 0.9811207671561283
Recall: 0.5001527650473572
F1 Score: 0.6625518567236669
ROC AUC: 0.7489415336567261
```



2.2 Random Forest

Random Forest is a widely utilized machine learning algorithm known for its high predictive accuracy. It belongs to the ensemble learning family and is effective for both classification and regression tasks.

The Random Forest algorithm operates in the following manner. Firstly, an initial prediction is established. In classification, this is typically set as the initial predicted probability of 0.5, while in regression, it involves using the mean of the training samples as the initial prediction. Subsequently, a collection of decision trees is constructed based on the residuals from the previous prediction. Each subsequent prediction is obtained by aggregating the output of the previous step with the output of the newly built tree. The output of each tree is scaled by the learning rate before being combined. As a result, the final output is generated by summing the initial prediction with the scaled outputs of all the trees. A visual depiction of the Random Forest algorithm is presented below.



(Cutler, Adele, Cutler, David, & Stevens, 2011)

2.2.1 Approach

In this project, the Random Forest algorithm was utilized for the prediction of 30-day hospital readmissions. In order to tackle the issue of high dimensionality in the data, a feature reduction technique was implemented. Instead of using PCA (Principal Component Analysis) due to processing limitations for large datasets, the approach involved transforming the data into a 1D array using means, standard deviations, maximums, minimums, and differences as features. This method effectively reduced the dimensionality of the data while considering important statistical measures of the variables.

To overcome the issue of class imbalance, the SMOTE technique was utilized. SMOTE generates artificial samples for the minority class, effectively balancing the distribution of classes and providing more representative training data. This strategy effectively mitigated the impact of class imbalance on the performance of the model.

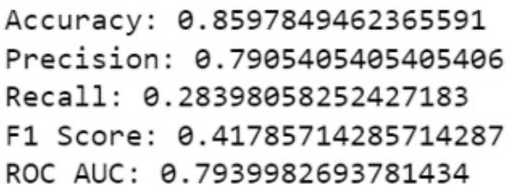
The model was constructed with 400 estimators. Increasing the number of estimators has been shown to enhance the model's performance by mitigating overfitting and improving overall stability. Additionally, the random_state parameter was set to 42 to ensure reproducibility. By fixing the random state, consistent and comparable results could be obtained when executing the code multiple times.

2.2.2 Result

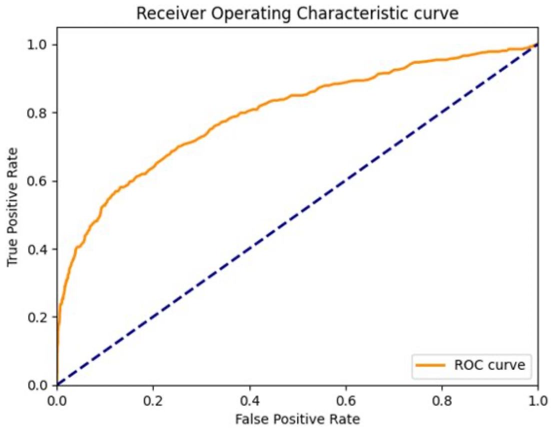
The research project aimed to predict 30-day hospital readmissions, with the goal of improving patient care and enhancing the efficiency of the healthcare system. The results achieved using the Random Forest model were highly encouraging.

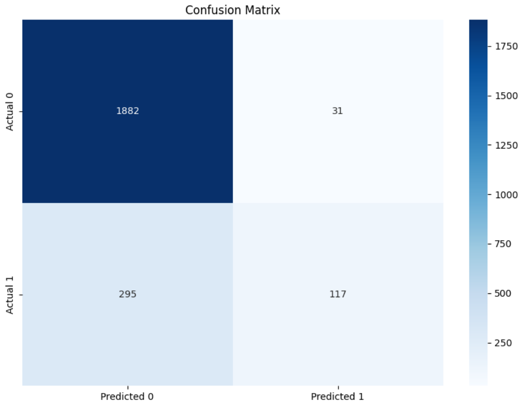
The model demonstrated a notable accuracy of 0.8598, suggesting a high level of correctness in predicting whether a patient would be readmitted within 30 days. With a precision score of 0.7905, the model showcased its ability to accurately identify positive instances, further contributing to effective patient care management.

Moreover, the ROC AUC score of 0.7940 indicated the model's strong discriminatory power in distinguishing between readmission cases and non-readmission cases. This implies that the model performed well in correctly classifying patients and minimizing false positives.



```
Accuracy: 0.8597849462365591
Precision: 0.7905405405405406
Recall: 0.28398058252427183
F1 Score: 0.41785714285714287
ROC AUC: 0.7939982693781434
```





From the confusion matrix above, the model has excellent ability in classifying negative data. We can tell by the low False Positive Rate (FPR) at 1.62%. However, the model did not perform as good in classifying positive data, with False Negative Rate (FNR) at 71.6%.

3. Finding

Our exploration of various model tuning methods aimed at improving performance. We focused primarily on feature selection techniques and oversampling techniques like Synthetic Minority Over-sampling Technique (SMOTE). However, our experiments revealed that these methods might not significantly enhance the prediction accuracy of the models.

3.1 Feature Selection Techniques

Our initial investigation involved feature selection techniques. Feature selection is typically beneficial when data exhibits strong linear relationships with the output. However, this may not be as effective for models like random forest and gradient boosting because of their inherent ability to handle non-linear relationships and interactions between features.

Our random forest model, combine groups of decision trees, have the capability of capturing nonlinear relationships between features and the output. This is especially relevant when the dataset includes diverse types of patient information with complex relationships. Consequently, feature selection methods may not significantly enhance the accuracy of random forest models.

Comparable to random forest model, our gradient boosting model also manage non-linear relationships and interactions between features. The model combines weak models sequentially to optimize a loss function, already consider feature importance during training. This consideration makes feature selection less necessary for gradient boosting models, as our experiments confirmed.

3.2 Oversampling Techniques: SMOTE

Our exploration extended to oversampling techniques, specifically SMOTE, to address class imbalance. However, the technique was not effective in increasing accuracy. This outcome might be because both our models, to some extent, have inherent mechanisms to handle class imbalance.

Our random forest model leverages the ensemble nature of multiple decision trees and diverse feature sets, which helps it manage class imbalance. Therefore, the application of SMOTE did not significantly enhance the performance of our random forest model.

Similarly, our gradient boosting model, specifically XGBoost algorithm that we used, assigns more importance to the minority class during the training process. This characteristic makes the model inherently capable of handling class imbalance, which may explain why the addition of SMOTE did not offer significant benefits to our gradient boosting model.

3.3 Summary and Future Directions

In summary, both feature selection and oversampling technique were not effective in improving the performance of the models. Since our models aim to achieve high accuracy for their respective purposes, we will focus on other areas of refinement to further enhance their performance. This investigation underscores the importance of understanding the characteristics and capabilities of the chosen models before employing certain tuning techniques.

4. Conclusion

In conclusion, our study compared the performance of Gradient Boosting and Random Forest models in predicting hospital readmissions. Both models produced encouraging results, with Gradient Boosting outperforming Random Forest in terms of precision and accuracy. Both models' ROC AUC ratings revealed their ability to differentiate between readmission and non-readmission instances.

While our efforts to improve performance through feature selection and oversampling techniques yielded no substantial gains, the models' overall performance was satisfactory. These results indicate that both the Gradient Boosting and Random Forest models have the potential to be useful tools for forecasting hospital readmissions and improving patient care management.

Further study might investigate alternative optimization strategies and evaluate the models' performance on diverse datasets and healthcare contexts in the future. We can make great progress in minimizing avoidable hospital readmissions, improving patient outcomes, and optimizing resource allocation in the healthcare system by continuously refining and upgrading these predictive models.

5. Reference

Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7_5. Download citation of Random Forests (researchgate.net)

Zhang, Tao & Lin, Wuyin & Vogelmann, Andrew & Zhang, Minghua & Xie, Shaocheng & Qin, Yi & Golaz, Jean-Christophe. (2021). Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning. Journal of Advances in Modeling Earth Systems. 13. 10.1029/2020MS002365.