Recognition Optimizing disease progression rates in Lou Gehrig Disease (ALS)

Yonghao Duan

1 Introduction

Amyotrophic lateral sclerosis(ALS) is a disease that degenerate nerve cells in the brain and spinal cord, which control voluntary muscle movement.[1] As ALS develops, patients will lose their ability gradually to control voluntary muscle movement. The severity of the disease is evaluated by a modified version of ALS Functional Rating Scale(ALSFRS-R) containing a list of 12 assessments regarding motor function, including speech, salivation, swallowing, handwriting, cutting (with and without Gastrostomy), dressing and hygiene, turning in bed, walking, climbing stairs, dyspnea, orthopnea, and respiratory insufficiency. Each measure ranges from 0 to 4, with 4 being the highest capability and 0 being no capability of function.
Predicting ALS progress of patients is hard due to the greatly heterogeneous performance for different patients. The goal of the project is to train machine models to predict disease progress of ALS patients based on the data collected in PRO-ACT database.[1]

2 Data Preprocessing

1. Data Collection

Overall, four datasets are used in the project. The first and second datasets are used for training models and the third and fourth dataset are used to test models. The first dataset used for the project is originally from Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database and then it was developed by Origent the sponsor company for this project. The PRO-ACT includes records of 10429 fully de-identified clinical patients with 74 features for each patient. The second dataset ALSFRS-R records diagnosis information over time after each patient's first diagnosis. It includes 66407 samples with 37 features.[1]

For the first dataset, the selected features for analysis are listed below in table 1. The column description provided details for each feature.

Table 1: Data Dictionary for Baseline Dataset

feature	structure	description
pid	id	patient ID, each patient has a unique ID
age_yr_bl	integer	age of this patient

f_male_bl	binary 0 and 1	0 represent female, while 1 represent male
race_bl	character	American Indian, Asian, Black/African American, Caucasian, and Other
ethnic_bl	character	Non-Hispanic or Latino, Hispanic or Latino
height_cm_bl	numeric	height of patient (in cm)
f_onset_bulbar_bl	binary 0 and 1	when first visit, if this patient has onset on bulbar diagnosis.
f_onset_limb_bl	binary 0 and 1	when first visit, if this patient has onset on limb diagnosis.
f_onset_spine_bl	binary 0 and 1	when first visit, if this patient has onset on spine diagnosis.
onset_days_bl	numeric (all negative)	disease onset day (to Time 0 - the start of the trial)
diag_days_bl	numeric (all negative)	clinical diagnosis day (to Time 0 - the start of the trial)
f_riluzole_bl	binary 0 and 1	when first visit, patient has taken riluzole (drug)
f_study_drug_bl	binary 0 and 1	when first visit, patient has taken study drug
q1_speech_bl	ordinary	0-4 levels. 4 means normal, 0 means lose speech ability
q2_salivation_bl	ordinary	0-4 levels. 4 means normal, 0 means lose salivation ability
q3_swallowing_bl	ordinary	0-4 levels. 4 means normal, 0 means lose swallowing ability
q4_handwriting_bl	ordinary	0-4 levels. 4 means normal, 0 means lose handwriting ability
q5_cutting_bl	ordinary	0-4 levels. 4 means normal, 0 means lose cutting ability
q6_dressing_and_hygiene_bl	ordinary	0-4 levels. 4 means normal, 0 means lose dressing ability
q7_turning_in_bed_bl	ordinary	0-4 levels. 4 means normal, 0 means lose turning ability
q8_walking_bl	ordinary	0-4 levels. 4 means normal, 0 means lose walking ability
q9_climbing_stairs_bl	ordinary	0-4 levels. 4 means normal, 0 means lose climbing ability
q10_respiratory_bl	ordinary	0-4 levels. 4 means normal, 0 means lose respiratory ability
alsfrs_total_bl	numeric	sum of above 10 levels. range from 0-40
alsfrs_r_total_bl	numeric	sum of q1-q9 + r1-r3
r1_dyspnea_bl	ordinary	a new sub question from q10.
r2_orthopnea_bl	ordinal	a new sub question from q10.
r3_respiratory_insufficiency_bl	ordinal	a new sub question from q10.
face_subscore_bl	ordinal	subscore on facial ability, range 0-12
hand_subscore_bl	ordinal	subscore on hand ability, range 0-12
body_subscore_bl	ordinal	subscore on body ability, range 0-12

leg_subscore_bl	ordinal	subscore on leg ability, range 0-12
chest_subscore_bl	ordinal	subscore on chest ability, range 0-12
fine_subscore_bl	ordinal	subscore on fine ability, range 0-8 ?
gross_subscore_bl	ordinal	subscore on gross ability, range 0-12
alsfrs_total_comb_bl	numeric	sum of above 10 levels. range from 0-40
alsfrs_r_total_comb_bl(y)	numeric	sum of q1-q9 + r1-r3
alsfrs_r_total_derived_bl	numeric	if r1-r3 are empty, then use the same value as q10. q1-q9+3*q10. if r1-r3 not empty, then sum of q1-q9 + r1-r3
f_gast_bl	binary(0,1)	with or without gastrostony
vc_liters_bl	numeric	Vital Capacit, unite of liters
creatinine_bl	numeric	a breakdown product of creatine phosphate in muscle
phosphorus_bl	numeric	Normal ranges are 1-1.5 mmol/L
uric_acid_bl	numeric	urine acid
sbp_mmhg_bl	numeric	Systolic Blood Pressure
dbp_mmhg_bl	numeric	Diastolic Blood Pressure
pulse_bpm_bl	numeric	beats per minute
resp_rate_bl	numeric	respiratory rate
temp_C_bl	numeric	body temperature
weight_kg_bl	numeric	baseline weight
exp_vc_bl	numeric	matched by gender, age and height
pexp_vc_bl	numeric	vc_liters_bl/exp_vc_bl
bmi_bl	numeric	body mass index(weight(kg/height(m) * height(m))

For the second dataset, we used features including time(t), pid and ALSFRS_R score. Time(t) describes days since the patient has been confirmed with ALS disease. *Pid* refers to a unique value for each patient. *ALSFRS_R score* is the total score of 12 assessments regarding motor function and it is also the target value of model used in this project.[2]
For each patient, only two features changes across time: "t" and prediction results.Therefore, we addressed the analysis on feature "t" and extract time information for building models. We thus created new features based on t and build related models to extract relationships between feature "t" and other relevant features.

2. Data Preparation

We conducted four phases of data preparation:

1) Data Merging: The first and second datasets are merged by the feature *pid*, as the single feature shared by two datasets. We deleted the samples that do not involve information.

2) Missing Data Cleaning: a) We deleted any feature that contains over 20 percent missing values. b) Other missing values of continuous variables are imputed with the mean of the variable and missing values of categorical variables are imputed with the mode of

the variable. By imputing data to fill the missing values, we aim to achieve best accuracy in modeling.

3) We transformed categorical variables into one-hot encoder so that machine learning algorithms can deal with data in number format rather than string format. The external dataset is processed with the same procedures.

3 Exploratory Analysis

Figure 1 below shows how six patients' performances change with time. It consolidated the evidence that ALS disease is heterogeneous and the prediction is hard to achieve a great accuracy. The X axis represents month while the Y axis represents the label score. For the patients 1-3 patients, while they shared similar initial scores, their disease progress differently. For patients 4-6, the pattern is reversed. They started with different initial scores but progressed with similar trends. For each patient with different time t, only feature "t" and prediction results change. So feature "t" is crucial and we extracted rich information for models. So, we created new features related with t and select models, which can extract relationships between feature "t" and other features.

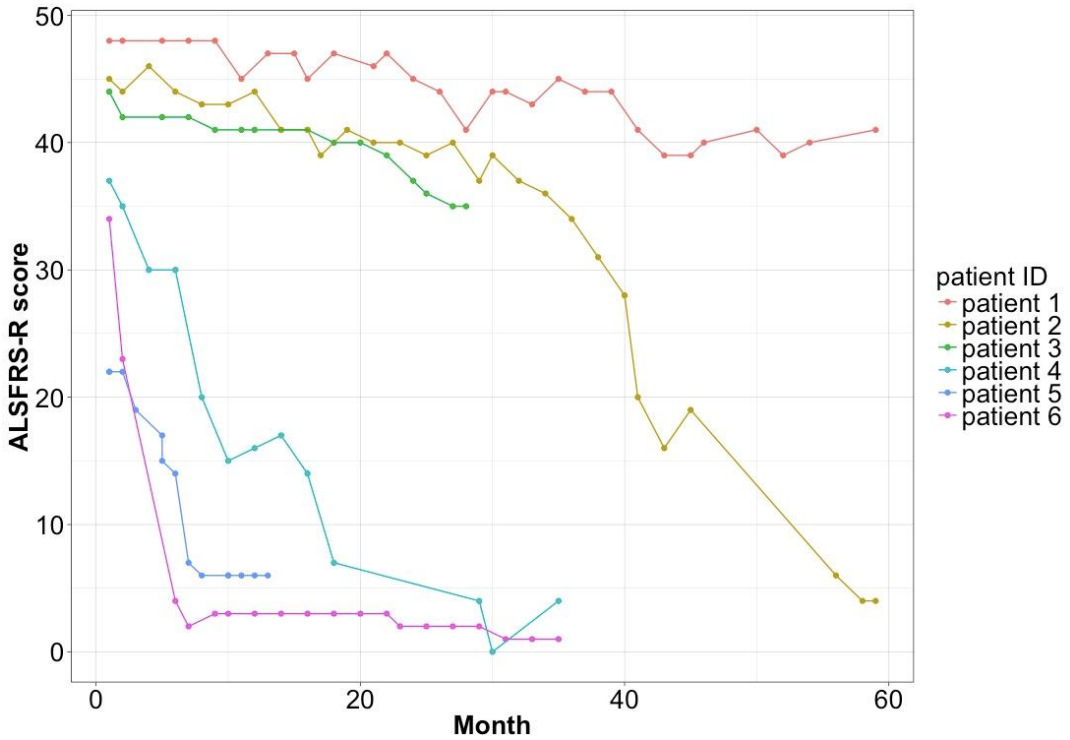

Figure 1: Six Cases of Patient's Muscle Performance through Time

Figure 2 shows a distribution of ALSFRS-R score respect to month. Visit frequency decreases with time. After month 40, there are less than 10 observations per month. Because most people will die gradually after patients were confirmed with ALS disease.

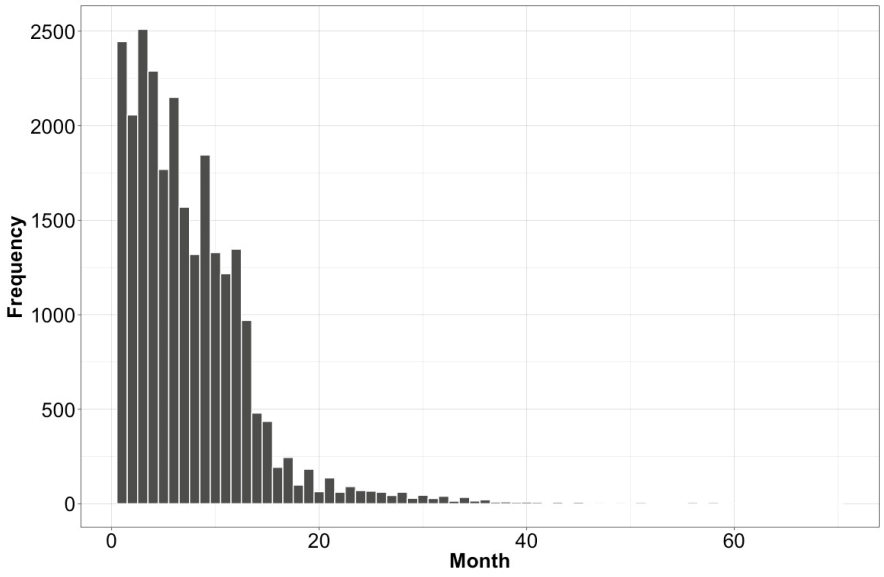

Figure 2: Visit Frequency within Each Time Period.

Figure 3 shows distributions of ALSRS_R score for different months. It is from month 8 to month 20. These distributions indicated high variance and confirmed our hypothesis that death typically occurs within 3 - 5 years of diagnosis.

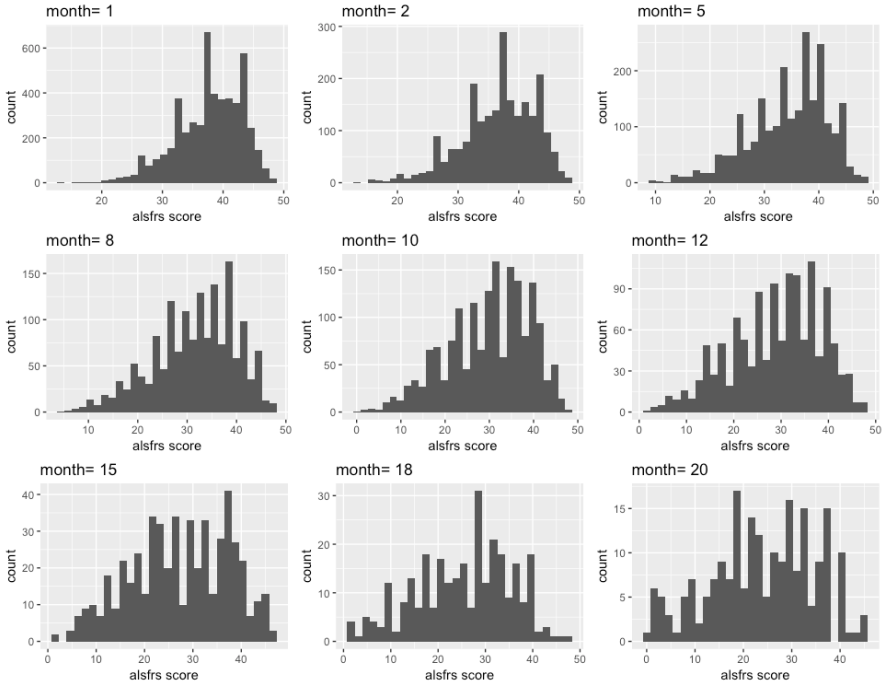

Figure 3: ALSFRS_R Score Distribution Within some time Period

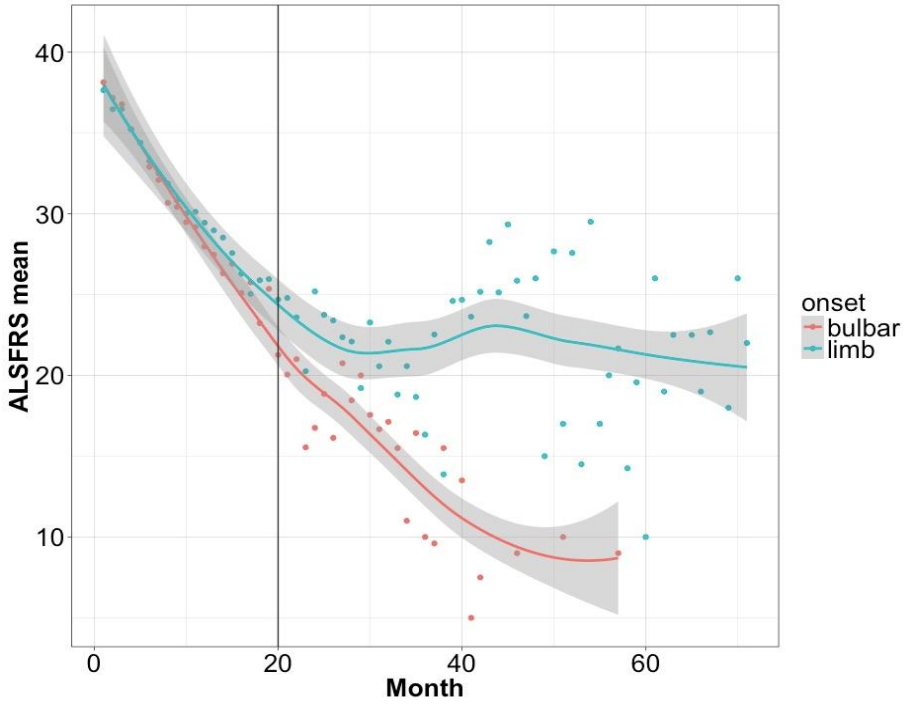

Figure 4: Onset Performance

Figure 4 shows a distribution between limb and bulbar, which are onset sections where the disease started firstly. It showed that the disease progress started to diverge after 20 months because of different onset sections. This feature can be used for model selection.

4 Model

1. Model selection

The above exploratory data analysis yielded three characteristics about the datasets: 1) It is mixed data that combines categorical and continuous data type; 2) A large amount of missing values are included in the original dataset; 3) Outliers are such as patients living over five years are included in the datasets. Taking these characteristics into considerations, we decided the model selection methods, as described below.

Table 2: Some Characteristics of Different Learning Methods [5]

TABLE 10.1. *Some characteristics of different learning methods. Key:* ▲*= good,* ◆*=fair, and* ▼*=poor.*

Characteristic	Neural Nets	SVM	Trees	MARS	k-NN, Kernels
Natural handling of data of "mixed" type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▼	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▼	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▼	◆
Interpretability	▼	▼	◆	▲	▼
Predictive power	▲	▲	▼	◆	▲

As the table showed above, the green triangle represents a good model to fit the characteristics listed in the left column. For tree model specifically listed in the table, it performs well for dealing with datasets with the first six problems, especially for mixed data, missing values and outliers. Given the advantages of tree models and the characteristics of our dataset, we primarily select tree models to perform modeling. However, tree-based model cannot perform well in extracting linear combinations of features. Considering the ability extracting linear combinations of features, neural network is suitable to address such problems. Therefore, we selected tree-based models and the neural network model for the regression analysis. We expected these two algorithms could mutually complement each other and generate a better prediction result. We also extracted the relationships between feature "t" and other features as the importance of time in our prediction.

2. Tree-based model

The figure 5 shows how one single decision tree split features. Decision tree is to divide features into several small parts based on Gini index. It indicated that the first important feature is baseline score and the second important feature is time t. The baseline score and feature t are the most relevant factors with predicted label. The split result echoes with the result shown in exploratory data analysis.

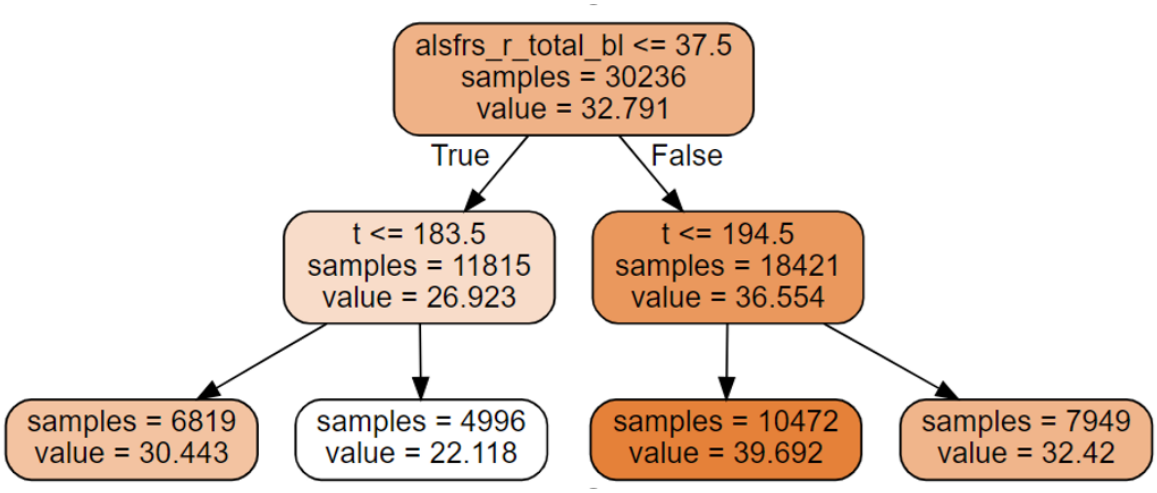

Figure 5: Single Decision Tree

The two tree-based models used for modeling are boostong methods: XGBoost and LightGBM. Boosting is an ensemble method that creates a strong model based on ensemble of weak classifiers, according to how correlated the learners are to the actual target variable. The errors of the previous model are corrected by the next predictor, by adding models on top of each other iteratively until the training data is accurately predicted or a maximum number of models are added. This process consists three steps: (1) an initial model F_0 is defined to predict target variable y. (2) a new model h1(one single decision tree) is fit to the residuals ($y-F_0$) from last step (3) F1, the boosted version of F0 is generated by combining F0 and h1. Therefore, the mean squared error from F1 is lower than that from F_0: $F_1(x) <- F_0(x) + h1(x)$. This can be repeated m iterations until residuals are maximumly minimized: $F_m(x) <- F_{m-1}(x) + h_m(x)$.

For XGBoost, it follows the principle of gradient boosting and uses second-order derivation for loss function to obtain better accuracy [3]. For LightGBM, it produces much more complex trees by following leaf wise split approach rather than a level-wise approach.[4]

3. Deep Neural Network

We built neural networks with six hidden layers. The neural network is fully connected and it means that every neuron in each layer will connect every neuron in other layers. The number of neurons of input layer is the same as the number of features. The output layer has one neuron, the predicted value. Based on neurons between the input layer and the first hidden layer, it demonstrates that all of features will be combined in linear combinations. That's why we selected neural network for modeling.

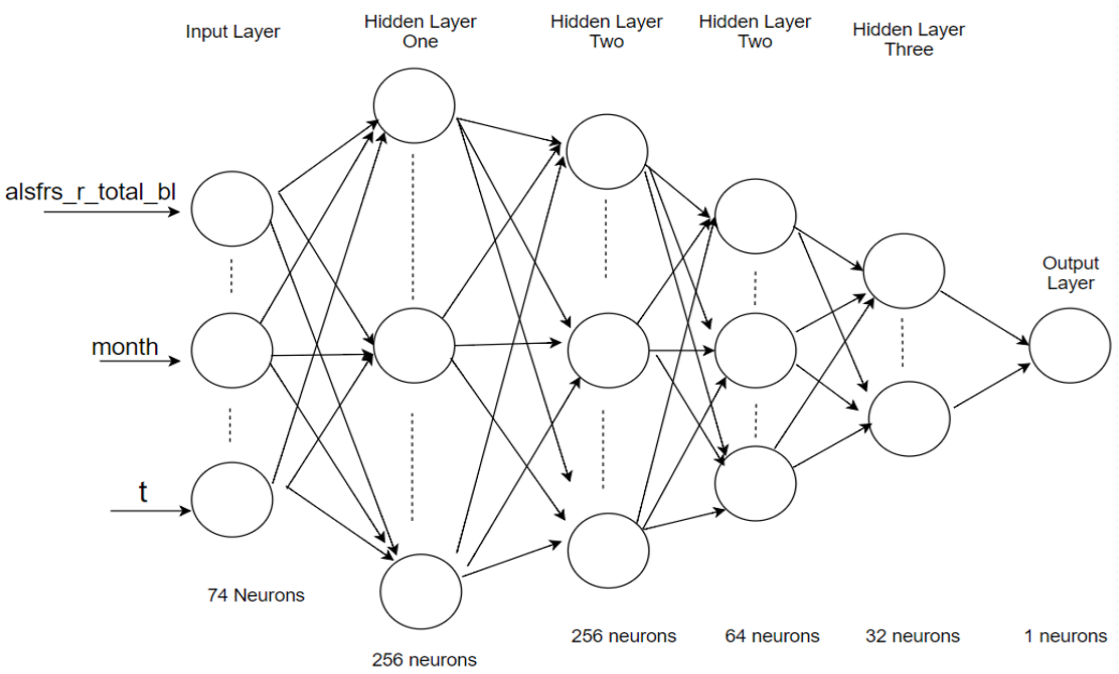

Figure7: Deep Neural Network

4. Weighted Ensemble

Weighted Ensemble method is to combine several machine learning models into one predictive model in order to decrease variance, bias and improve prediction ability. This approach allows the production of better prediction results by complementing advantage of different models.

Firstly, each algorithm will be trained separately by using 10-fold cross validation. The 10-fold cross validation means that data will be split into 10 groups where 9 groups will be used as train dataset and one left group will be used as valid dataset. The cross validation is used to estimate models on unseen data.

Secondly, we used weighted ratio, depended on best training accuracy, to assign different ratios for the prediction result from each model. Figure 6 shows full steps for weighted ensemble.

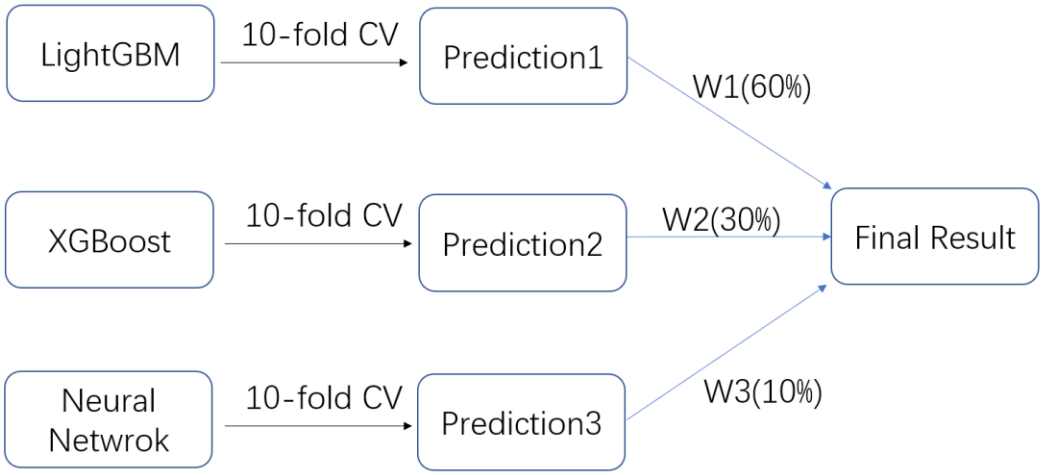

Figure 6: Weighted Ensemble

5. Performance and Result

1. Model Evaluation Criteria

The criteria used to evaluate models contains R2, RMSE, slope, intercept and skewness. R2 is correlation coefficient ranging from 0 to 1 and it describes percentage of the overall variance explained by model. RMSE is the square root of the mean squared error to describe remaining measurement variance not explained by model. A model with RMSE in a range of 10-15% is regarded as good model. The slope and intercept are calculated by using predicted value and true value. skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean.

2. Internal Validation.

Table 3 shows the model performance for the internal dataset. From the table, tree-based models (XGB and LGB) perform well while DNN generate less desirable results. When we combined three tree models in different ratios, the weighted ensemble method improves on R2 and RMSE. The best R2 is 0.706 and RMSE is 4.864 accordingly

Table 3: Comparisons of different models

	XGB	LGB	DNN	Weighted Ensemble
R2	0.702	0.704	0.606	0.706
RMSE	4.896	4.882	5.631	4.864
Slope	0.988	1.001	0.946	1.009
Intercept	0.460	0.025	2.246	-0.194
Skewness	-0.537	-0.522	2.194	-0.468

Moreover, the prediction result of the first year is shown below. From what we have learnt from figure 2, most of data falls within the first 20 months, indicating the result of the first year is important to patients and doctors. The improved percentage is shown in the table 4 and all of metrics has been improved by using the weighted ensemble method.

Table 4: Comparisons of different models' performance in the first year's data

	XGB	LGB	DNN	Weighted Ensemble
R2	0.724(↑3.13%)	0.725(↑2.98%)	0.589(↓2.81%)	0.727(↑2.97%)
RMSE	4.214(↓13.93%)	4.206(↓13.85%)	5.140(↑8.72%)	4.190(↓13.86%)
Slope	0.991	1.000	0.874	1.005
Intercept	0.337	0.051	4.902	-0.049
Skewness	-0.538	-0.556	2.769	-0.488

3. Bias Analysis

Mean Prediction Error (MPE) is used to measure the bias in prediction results, which is different between true values and predicted values. In order to assess the bias by month, bootstrap method is used to generate frequency plot of MPE. We bootstrap samples 10000 times for each month. The figure shown below is MPE of month one. The zero line is in the middle of the frequency plot, fairly enough to show the prediction for month 1 has no bias.

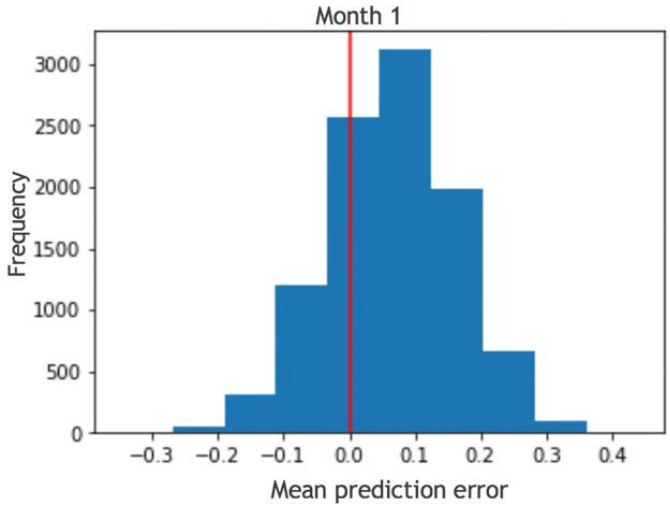

Figure 7: MPE for Month 1

4. External Validation

The external dataset is used to test model prediction performance and it is preprocessed with same procedures as the internal dataset. The comparison results between two datasets are shown below. The R2 and RMSE results for two datasets are closed to each other, meaning a model has good generalization ability

Table 5: Comparisons of internal validation and external validation

	Internal Validation (sub-dataset: 15 months)	External Validation (full data: 15 months)
N	28127	1538
R2	0.713	0.717
RMSE	4.486	4.560

6. Conclusion

We selected models given the characteristics of datasets and the only changeable feature t for each patient. This drives us to use tree-based models and the neural network model. These two types of models were combined to complement each other so as to achieve best prediction accuracy. The result yielded the best performance for R2 is 0.706 and for RMSE is 4.864 during the whole period and the best performance achieves 0.727 for R2 and 4.190 for RMSE in the first one year. The model performs even

better in external dataset and it means that model maintains a good generalization capability. What is more, there is no strong bias across all models for the first one year data. Besides, a python package is built to evaluate the performance among different models.

7. CONTRIBUTIONS

My contribution to this project is mainly on data preprocessing and modeling part. The modeling part contains training LightGBM, XGBoost and DNN models.

References

[1] Pooled Resource Open-Access ALS Clinical Trials Database, https://nctu.partners.org/ProACT/Document/DisplayLatest/2

[2] General use ALS models: Validation report, Origent

[3] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[J]. 2016:785-794.

[4] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu: LightGBM: A Highly Efficient Gradient Boosting Decision Tree

[5] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer.