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

He Huang
University of Rochester

Introduction

Amyotrophic lateral sclerosis (ALS), also known as motor neuron disease (MND) or Lou Gehrig's disease, is a specific disease which causes the death of neurons controlling voluntary muscles. We are trying to create a series of models that predict ALSFRS-R scores that evaluates the progression of the disease over time with PRO-ACT database.

In PRO-ACT, these records are all patients getting ALS disease. Some also use the term motor neuron disease for a group of conditions of which ALS is the most common. It may begin with weakness in the arms or legs, which is limb onset. It may begin with difficulty speaking or swallowing, which is bulbar onset. About half of people develop at least mild difficulties with thinking and behavior and most people experience pain. Most eventually lose the ability to walk, use their hands, speak, swallow, and breathe.

The cause is not known in 90% to 95% of cases, but is believed to involve both genetic and environmental factors. The remaining 5–10% of cases are inherited from a person's parents. About half of these genetic cases are due to one of two specific genes. The underlying mechanism involves damage to both upper and lower motor neurons. The diagnosis is based on a person's signs and symptoms, with testing done to rule out other potential causes.

Data description

Data was downloaded from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database. From that database, we find that there are over 10429 ALS patients who involved in industry clinical tests. PRO-ACT was made by multiple trials so that different patients can have different types of information.

There are 10 aspects of activities measured in these trails which contains speech, salivation, swallowing, handwriting, cutting food and handling utensils (with or without gastrostomy) dressing and hygiene, turning in bed and adjusting bed clothes, walking, climbing stairs, breathing. The total ALSFRS-R score is obtained by 12 questions answered by patients when they come to the hospital. And these 12 questions are used to evaluate the condition of patients.

There are 75 features and 10429 patients in ur_adbl_proact.csv with ‘pid’ indicates different kinds of people. This dataset is regarded as baseline data. Also, there are 38 columns and 66407 rows in ur_adep_proact.csv, which is seen as endpoint data. So our goal is to develop an algorithm that uses the baseline data of patients to predict ALSFRS-R score of patients in the future.

There are far more records in ur_adep_proact.csv than in ur_adbl_proact.csv because duplicate records of one patient were recorded for multiple visit. All these features have different meanings. There are three features taken from two sheets since we think only three feathers are useful in data analysis: pid (patientID), t (time), r_alsfrs_r_total (ALSFRS score). The left features are not related to our response variable. The description of some important features is shown as below in Table.1. For these 12 questions, different numbers indicate different level of conditions. The higher the score, the better the body condition. The range of score is from 0 to 4. Our goal is to predict functional decline in ALS at a given time.

Data preprocessing

We use pandas to combine two sheets together: selecting the first sheet and choosing the first three features from the second sheet. Then we find that the merged data has 66407 rows and 77 columns. The only changeable feature in this file is t (time).

There are some missing values existing in these two sheets. If missed data is responsible variable, we delete the related records. If the percentile of missing values of a feature is more than 20%, we delete the feature. For discrete values, we fill in the blank with the value appearing most frequently in that column. Differently, we calculate the median of the entire data set to fill in the blank of continuous variable. After preprocessing, the dataset has its dimensions changed to 30236 rows and 74 columns. Features that contain highest 10 percentile of missing value are shown in Fig.1.

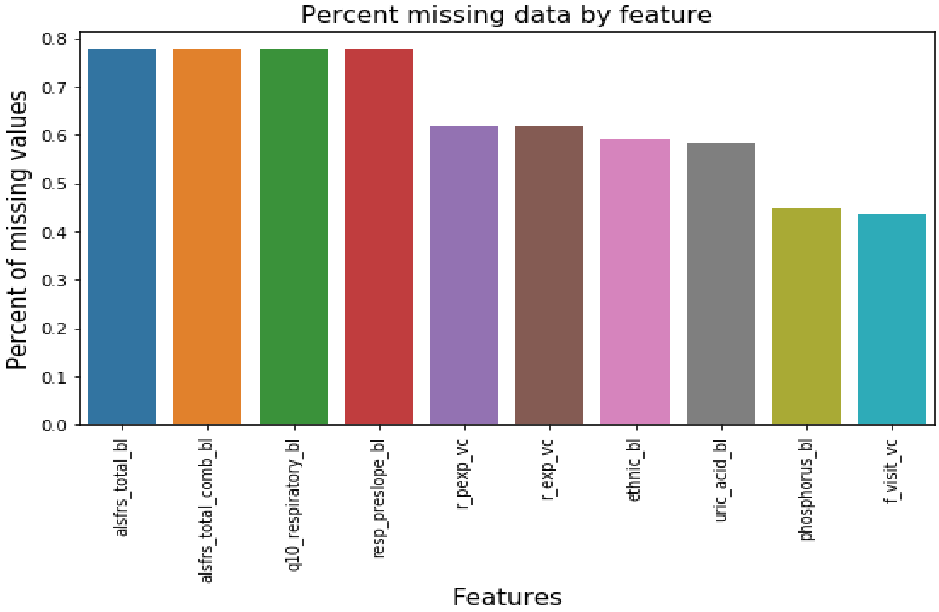


Fig. 1. Distribution of mean of ALSFRS total score with time

Feature name	Feature type	Feature description
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

Table.1. Description of some important features

However, when we pay attention to the dataset, we find that there are several columns are highly related to the alsfrs_total_comb_bl, so we need to discard these features. Alsfrs_r_total_comb_bl(y) is selected as the label. Then we randomly set 90% of data as training data, and 10% of data as testing data to use the 10-fold cross validation. Performance of the regression model was evaluated through internal 10-folad CV. For each patient in the data, the model was used to predict a patient's outcome value at future timepoints and was then compared to the patient's actual outcome at those times.

Exploratory analysis

First, we explore the relationship between time and response variable. ALS can be very heterogenous as is shown in Fig.2. We choose 60 mouths as time, and we choose ALSFRS-R score as y axis.

Patient1, patient2 and patient3 are top three lines which have a high muscle performance at baseline. However, patient1 just fluctuates among 60 months while patient2 decreases dramatically. Patient3 died at 30th month, far earlier than patient1 and patient2.

Patient4 and patient6 begins at month2, and ends at month38. Patient6 decreases rapidly at first few months, but it just remains stable in the next few months. Patient4 had his score decrease at a constant speed.

For patient5, his score begins at a low level, but it just lasts for only few months.

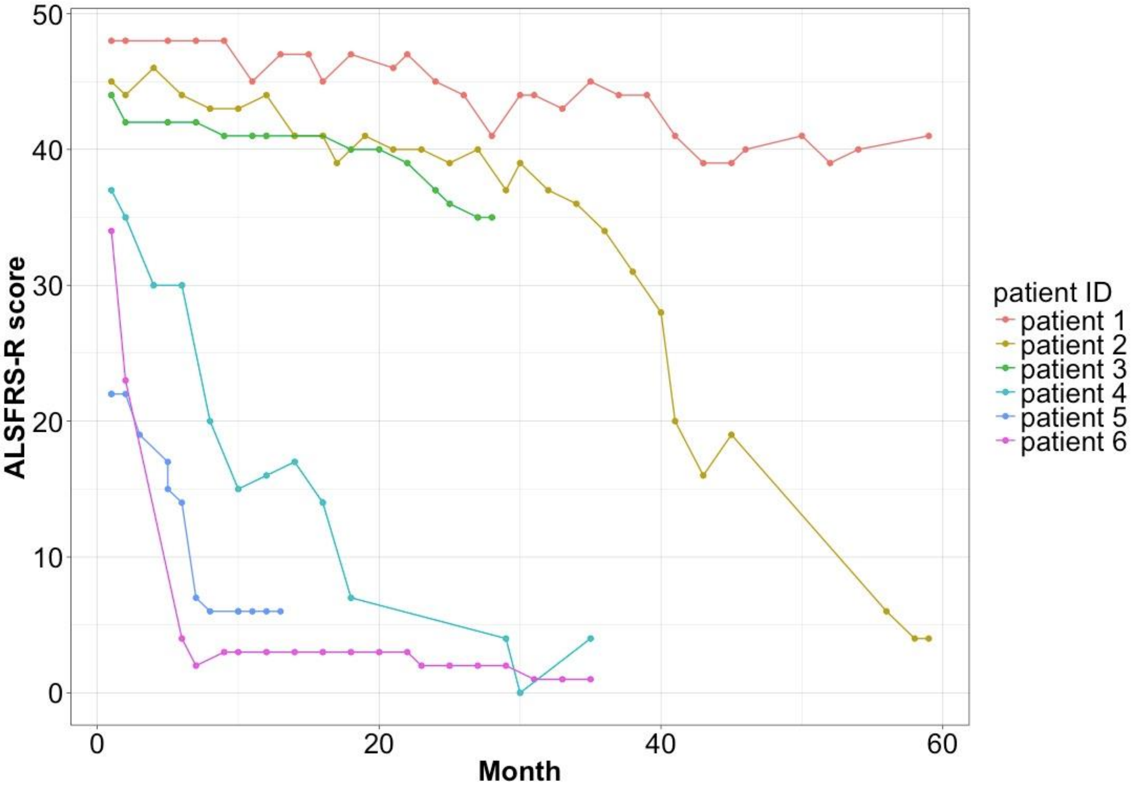


Fig.2. 6 cases of patient's muscle performance through time

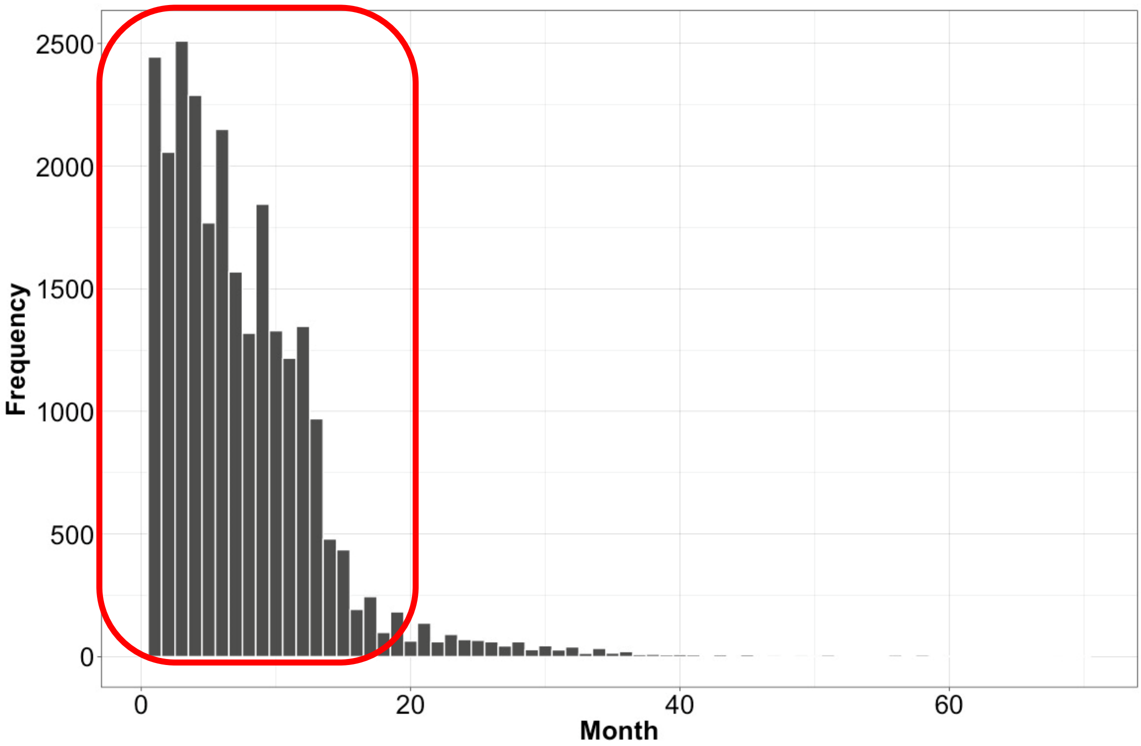


Fig.3. Frequency of visits in different time period

We have to calculate the frequency of the visits in every month. The Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) is chosen as the response variable. As shown in Fig.3, the frequency of patients' visit decreases remarkably along the time.

Afterwards, we calculate mean of response variable (ALSFRS_R) and plot it in Fig.3. However, data points begin to be more scattered after 20 months. Maybe during this process, some patients dead or some patients were new patients that their medical record did not cover over 20 months.

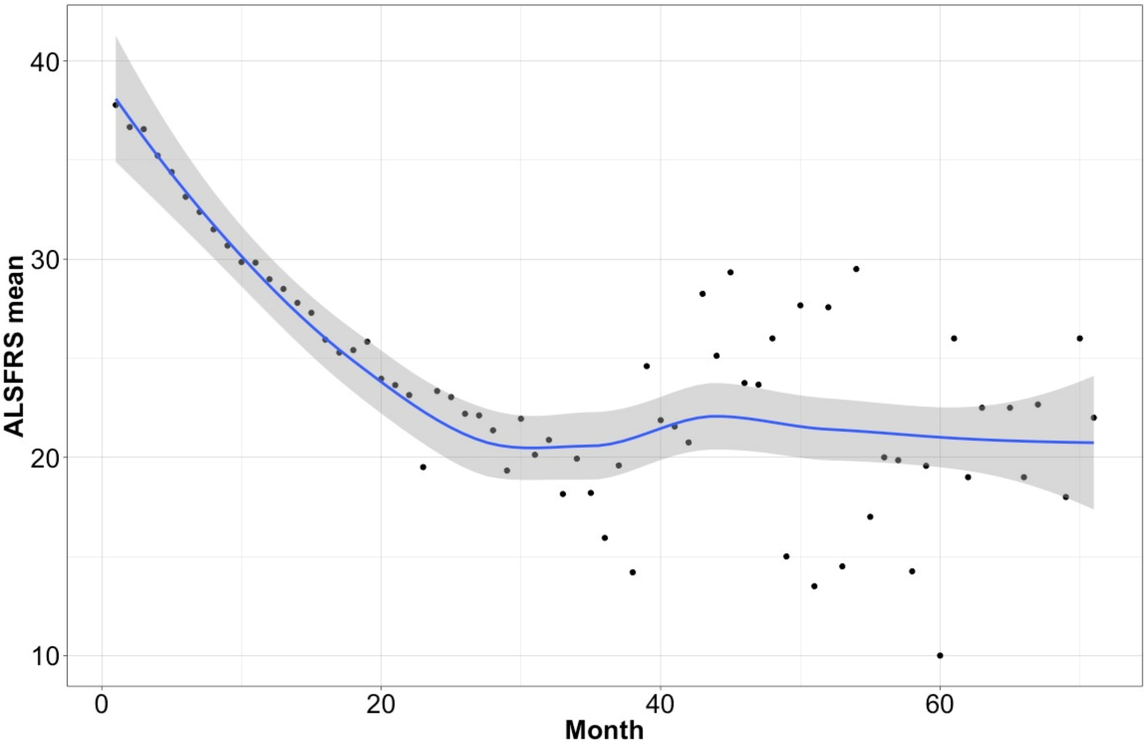


Fig.4. Mean ALSFRS_R of each month

Then we plot the distribution of gender, and onset in Fig.5 and Fig.6 respectively.

As we mentioned above, since lack of medical records, the plots will look like scatter plots after 20 months. For the distribution of gender, there is some slight difference exist in two genders. In Fig.6, people with bulbar onset has their scores decreases more rapidly than people with limb onset. Because people with bulbar onset has difficulty swallowing, eating and speaking, their bodies will become weaker than people with limb onset.

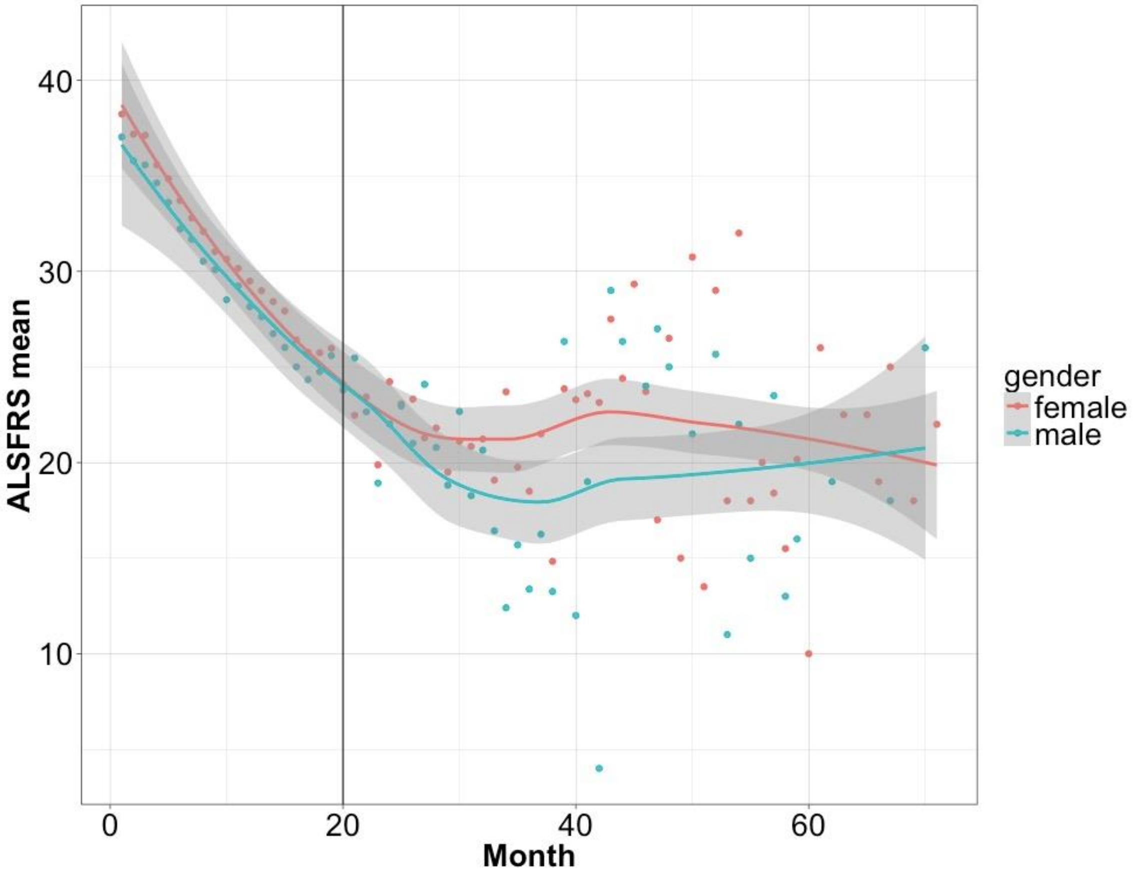


Fig.5. Gender distribution with month

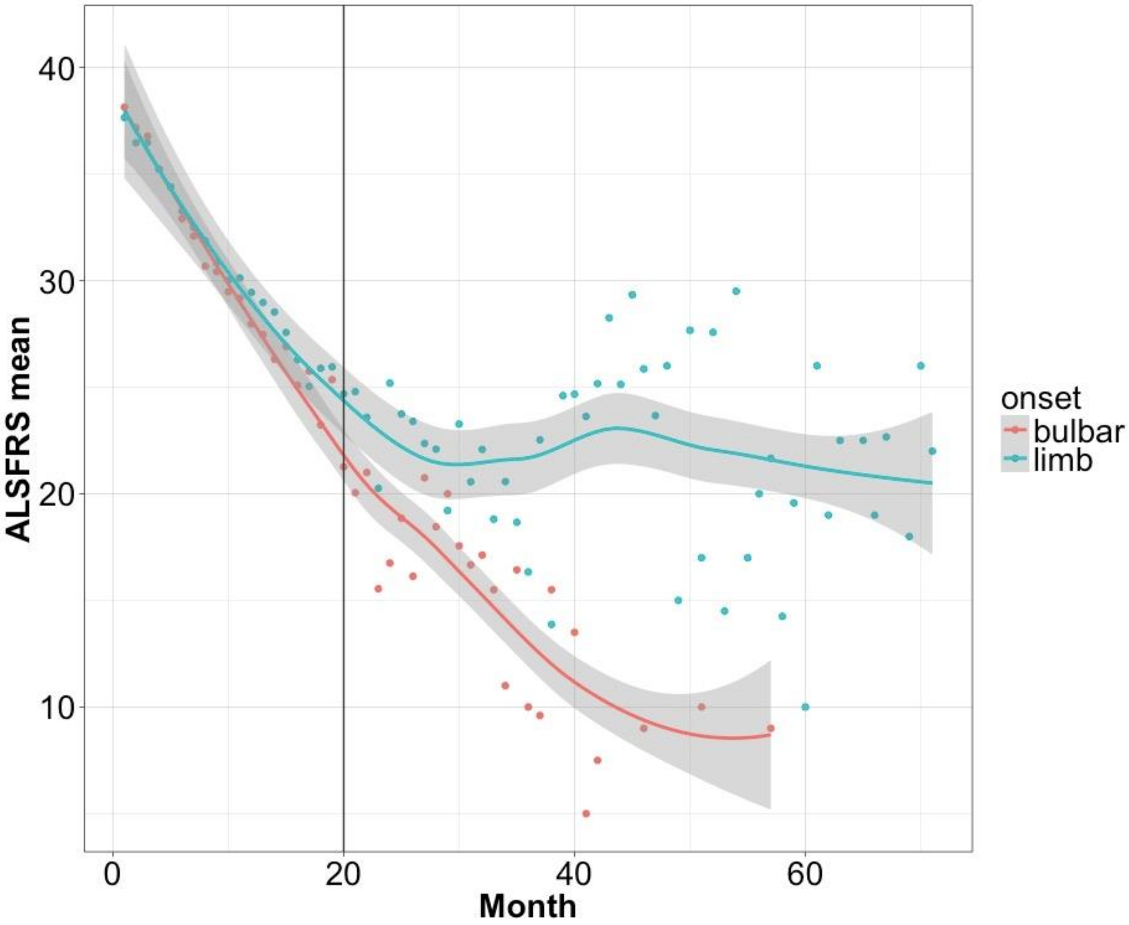
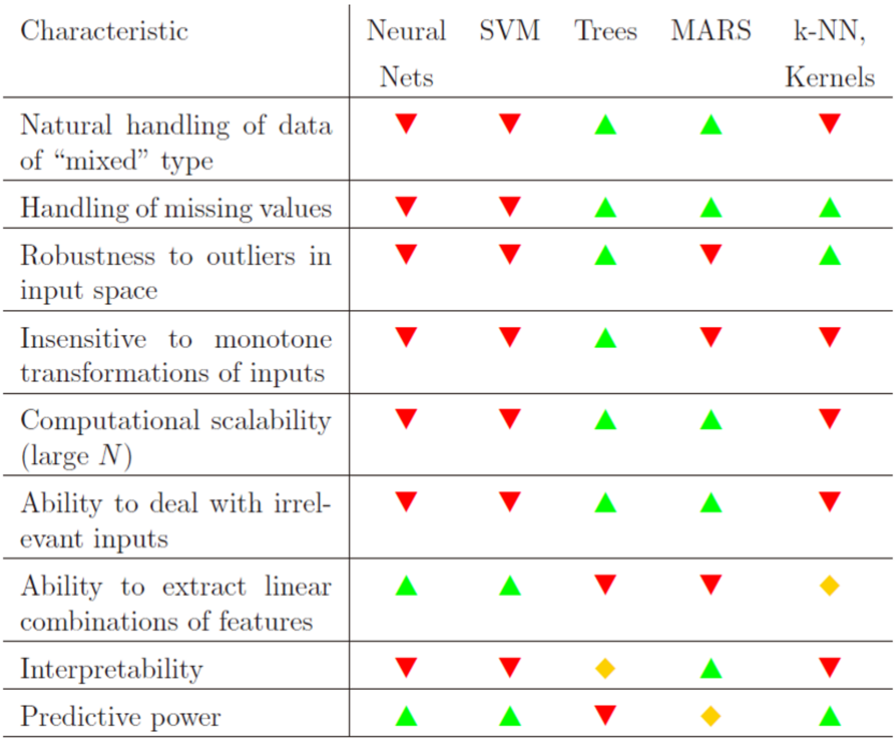


Fig.6. Age group distribution with month

Models and methods



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	▲	▲	▼	◆	▲

Table.2. Some characteristics of different learning methods

In Table.2, we compare power of different models, and we find tree models perform best among these models. But tree model has low predictive power and it cannot extract linear combinations of features.

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements machine learning algorithms under the Gradient Boosting framework. XGBoost provides a parallel tree boosting (also known as GBDT, GBM) that solve many data science problems in a fast and accurate way.

The same code runs on major distributed environment (Hadoop, SGE, MPI) and can solve problems beyond billions of examples. The individual machine learning model that form the ensemble are known as base learners and they are either from the same learning algorithm or different learning algorithms.

While XGBoost is similar to gradient boosting algorithm, it has some unique features which make it so interesting. XGBoost can penalize complex models through both L1 and L2 regularization to prevent overfitting. Also, XGBoost can effectively handle weighted data. Besides, XGBoost can make use of multiple cores on the CPU to allow parallel learning. Generally, compared to gradient boosting, XGBoost is faster and has a wider range of application.

NN performs well in learning from data. Artificial neural networks are one of the main tools used in machine learning. As the "neural" part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize.

Performance and result

We use correlation coefficient (R^2), root mean square error (RMSE), slope, intercept and skewness to evaluate the performance of each model.

The R^2 represents the proportion of the overall variance explained by the predictive model while RMSE is the measure for the remaining measurement variance not explained by the predictive model. The range of R^2 statistic is between 0 and 1. If $R^2 = 0$, the predictor model could not explain the measurement variance at all while $R^2 = 1$ illustrates that the predictor model fully explained the variance. It is interesting that a model with an RMSE that is approximately in the range of 10-15% of the outcome variable range yields useful applications empirically.

As is shown in Table.2, the tree models (e.g. XGB and LGB) perform well with the R^2 above 0.7 and RMSE of 4.8. However, DNN does not perform that good as tree models, with R^2 0.606 and RMSE 5.63. Then we combined three models together according with different weights, the weighted ensemble method made R^2 and RMSE increase to 0.706 and 4.864 respectively.

	XGB	**LGB**	**DNN**	**Weighted Ensemble**
R^2	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

Table.3. Comparisons of different models

The prediction of the first one year plays crucial rule to patients and doctors. Besides, most of the data shows a decreasing tendency in the first 20 months as we mentioned before. So, we want to find whether the models perform better in the first one year. Then we compare the performance of four models based on medical records in the first one year in Table.3. Clearly, all models except for DNN have improved a lot in R^2 and RMSE. Unfortunately, DNN performs even worse in the first one year than the whole period while the weighted ensemble is still the best one which has the highest R^2 and lowest RMSE with intercept close to 0 and the slope near 1. We can draw the conclusion that weighted ensemble is really a good improvement of individual models.

As showed in Fig6, features are input layers include: t, month, alsfrs-r-total-bl, gender and so on, while predicted ALSFRS-R score is output layer. Neurons are fully connected between each layer. Neural networks can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. As mentioned above that our dataset has only one changeable feature “t”. Therefore, detect relationship between t and other features is very important.

	XGB	**LGB**	**DNN**	**Weighted Ensemble**
R^2	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

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

We compare our result with previous work. The Fig.7 shows the R^2, the intercept and the slope from Origent's work: 0.696, 0.944 and 1.032. Our models have improved R^2, RMSE, intercept and slope.

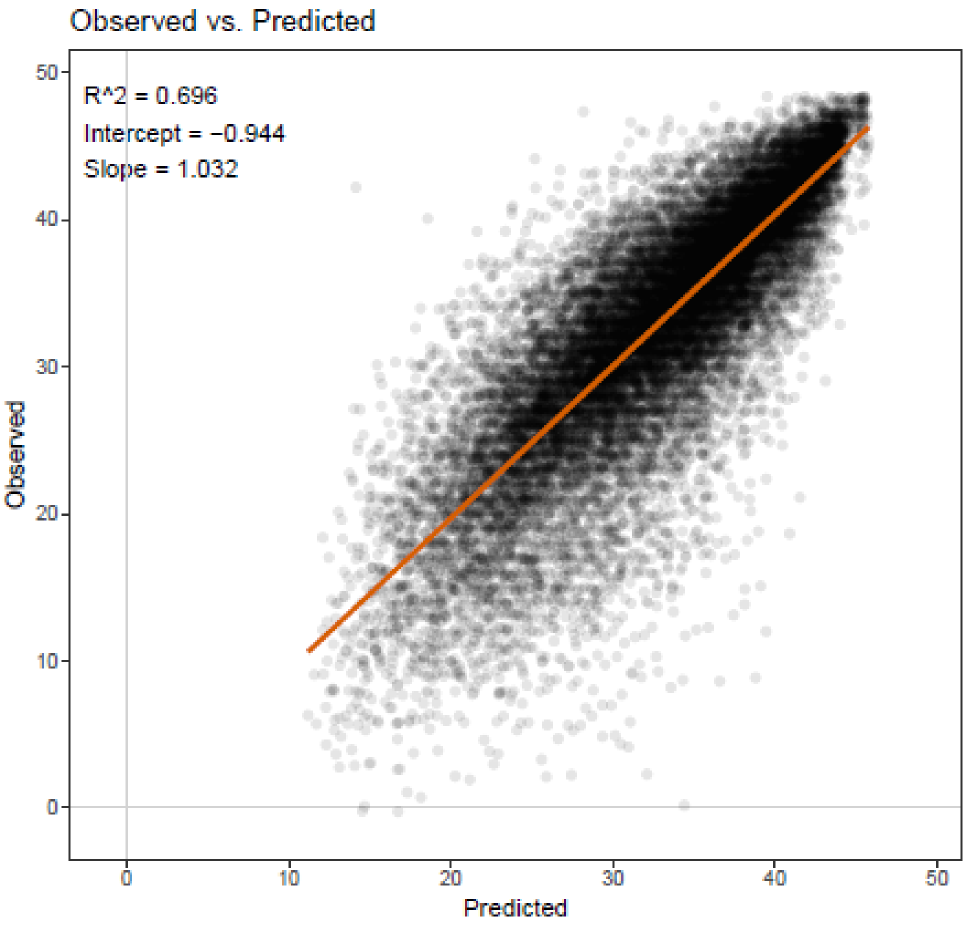


Fig.7. Results of Origent's work

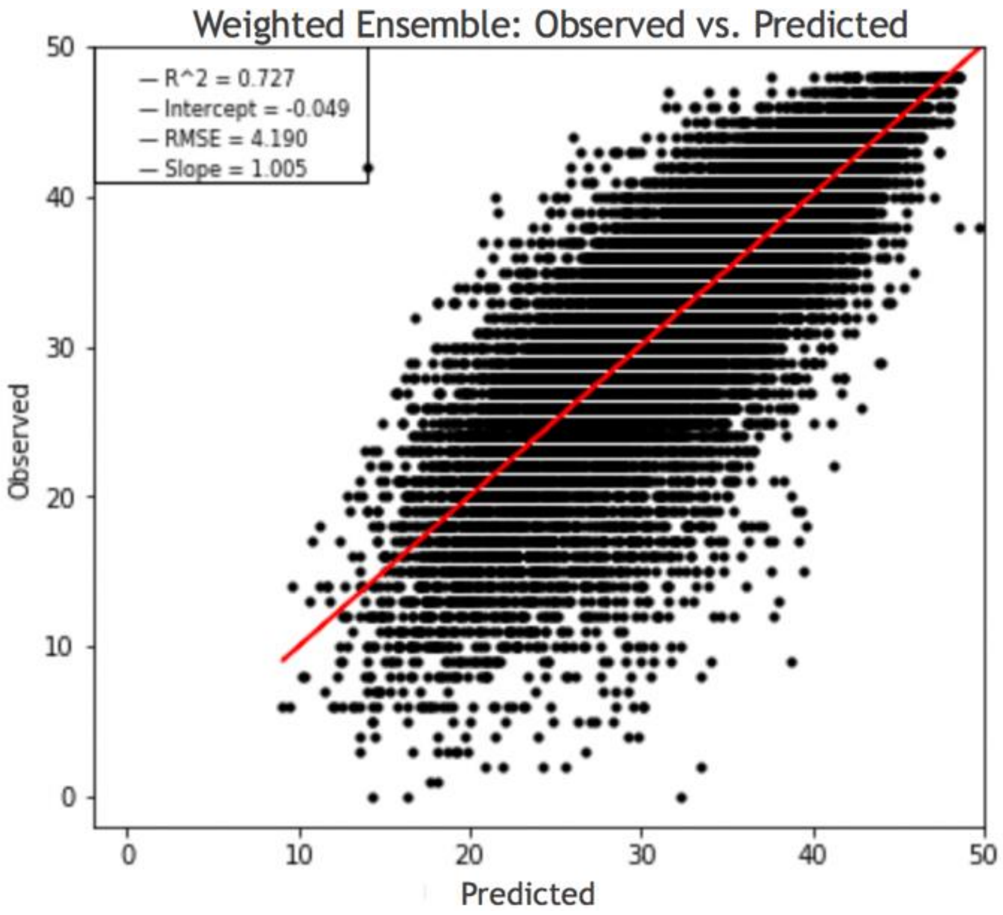


Fig.8. Results of weighted ensemble

Mean Prediction Error (MPE), which is defined as the difference between the observed value and predicted value, is used to measure the bias in results. The smaller the MPE, the better the result. To assess the bias by month, we bootstrap 10000 times for every month and get the frequency plot of MPE. If the zero line (MPE = 0) stands in 95% confidence interval, we can conclude that our result is convincing. For example, the zero line of Month one's frequency plot (Fig.9) is almost in the middle of the whole plot, thus we can state that the predictions for month 1 has no bias.

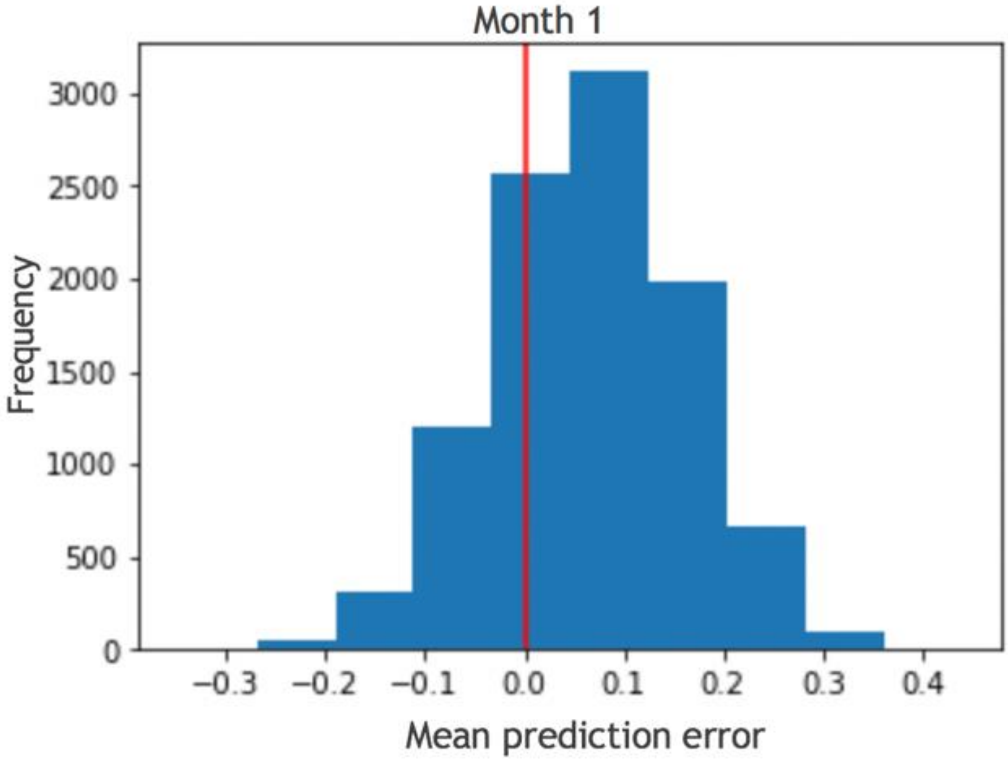


Fig.9. Zero line of mean prediction error for Month 1

We got the external test data from Origent after finishing the models. For the first 15 months, this test dataset has 1538 records. To compare the result in the same time range, we did the internal validation and external validation both for the first 15 months. For both internal and external validation, the results in Table.5 show that the R^2 and RMSE are similar in two validation process. Our models can be applied to any datasets since it performs really well.

	Internal Validation **(sub-dataset: 15 months)**	**External Validation** **(full data: 15 months)**
N	28127	1538
R^2	0.713	0.717
RMSE	4.486	4.560

Table.5. Comparisons of internal validation and external validation

Conclusions

In summary, we use different models to predict ALSFRS-R and perform their performance. We try different algorithms and models to analyze the problem. Neural network has the ability to extract linear combinations of different features. Although NN helps weighted ensemble preforms better, it is not the best model. So we try tree models and ensemble method to solve this problem. Comparing internal test with external test, result of the latter is even better, so we can conclude that models are not over-fitting. We do the experiments on only the first year's data points after we referring some clinical data, then we find that RMSE and R square become better. This is understandable because variance between patients is not very high in the first year.

Contributions

I finish part of Exploratory analysis and part of packages construction.

References

[1]. Blog, Guest. “Understanding the Math behind the XGBoost Algorithm.” Analytics Vidhya, 6 Sept. 2018, www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/.

[2]. "CatBoost vs. Light GBM vs. XGBoost." Towards Data Science, Towards Data Science, 13 Mar. 2018, towardsdatascience.com/catboost-vs-light-gbm-vs-xgboost-5f93620723db.

[3]. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning : with Applications in R. New York :Springer, 2013. Print.

[4]. Glander, Shirin. “Machine Learning Basics - Gradient Boosting & XGBoost.” Shirin's PlaygRound, 29 Nov. 2018, shirinsplayground.netlify.com/2018/11/ml-basics-gbm/.

[5]. "Interpretation Guides to Standardized Questionnaires Employed in the ALS CARE Database, Including the:" ALS SF-12, ALSFRS Interpretation Guides, www.outcomes-umassmed.org/als/sf12.aspx.

[6]. PRO-ACT. Pooled resource open-access als clinical trials database. https://nctu.partners.org/ProACT/ (Accessed February 21, 2015)