WIND SHEAR ESTIMATION

ANALYTICS ENGINEER CASE STUDY - FALL 2016

Abstract

The objective of this case study is to fulfill GE Analytic Engineer Certification Program. The current study provided detail analysis in three parts to estimate wind shear using wind load sensor data from field engineer. For labeled dataset, supervised learning models were adopted: General Linear Models (GLM), Classification and Regression Tree (CART), Neural Network (NN), Support Vector Machine (SVM), Random Forrest (RF) and eXtreme Gradient Boosting (XGB). Current case study showed satisfactory results with down-selected XGB model for both classification and regression. However, the author would NOT recommend GE Wind to launch the "Virtual Wind Shear Sensor" program at this time due to limited sample size, imbalanced dataset, neglected confounding factors, and missing risks and financial analysis. A pilot verification program is recommended for different wind turbine models at different sites.

Analytics Engineer Case Study - Wind Shear Estimation

September, 2016

Abstract

The objective of this case study is to fulfill GE Analytic Engineer Certification Program. In this case study, the wind load sensor data were provided by field engineer as a labeled dataset. The current study provided detail analysis in three parts to estimate wind shear. In Part A, classification models using wind turbine load sensor data (14 predictors) were explored to classify wind speed profile as Power Law, LLJ, Flat and Other. In Part B, regression models were developed to estimate wind shear (alpha) for Pow Law profile. Regression models for wind speed profiles of LLJ and Flat were studied in Part C.

For labeled dataset provided in the case study, supervised learning models (classification and regression) were considered, e.g. General Linear Models (GLM), Classification and Regression Tree (CART), Neural Network (NN), Support Vector Machine (SVM), Random Forrest (RF) and eXtreme Gradient Boosting (XGB). These models can be used for both classification and regression problems.

The dataset is split into training (75%) and test (25%) sets. 5-repeat of 10-fold Cross-Validation (CV) was used for model hyperparameter tuning with training data only. Model performance was evaluated with optimized hyperparameters using test data. Kappa and RMSE metrics were used to down-select classification model (Part A) and regression model (Part B & C) respectively. Current case study followed the common process in data science of Obtain, Scrub, Explore, Model and Interpret (OSEMN). The analysis program is R in RStudio environment.

Current case study showed **satisfactory results** with down-selected **XGB** model for both classification (Part A) and regression (Part B & C). However, the author would **NOT recommend** GE Wind to launch the "Virtual Wind Shear Sensor" program due to limited sample size, imbalanced dataset, neglected confounding factors, and missing risks and financial analysis. A **pilot verification** program is recommended for different wind turbine models at different sites.

Keywords: Analytic Engineer, Machine Learning, Supervised Learning, Classification, Regression, General Linear Model, Classification and Regression Tree, Neural Network, Support Vector Machine, Random Forest, eXtreme Gradient Boosting , Cross-Validation, Hyperparameter, Model Selection, Imbalance, SMOTE, ROSE, R, RStudio

1 Objective

The current case study was developed by Xu and Evans (2016) based on a dataset collected from one wind turbine with both wind speed measurements from a 5 sensor met masts and other parameters from load sensors. For every wind turbine, load sensors are installed at the hub to sense the nodding, pitching, yawing, and other loading parameters. Some of them are influenced by wind speed profile as shown in Figure 1. Wind speed profile also affects the wind turbine power output by changing the attack angle of wind on the turbine blades. The wind speed profile is characterized by wind shear, which is the gradient of wind speed (V) with respect to altitude (h). Wind speed profile can be measured by installing 5 wind speed (LIDAR) sensors at different altitudes on a costly met mast.

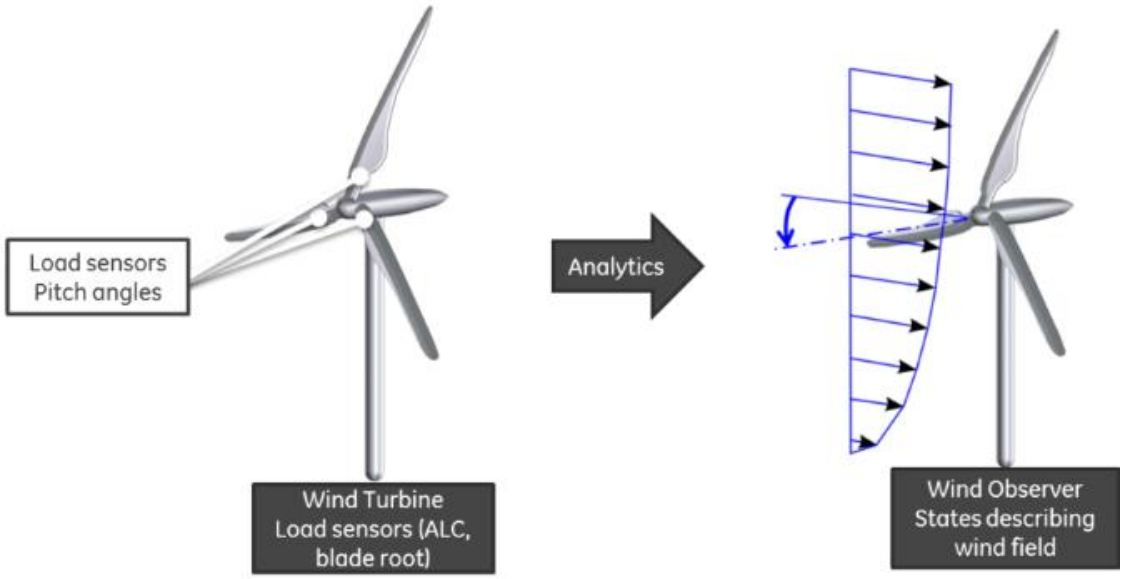


Figure 1 Illustration of using load sensor data to estimate wind shear

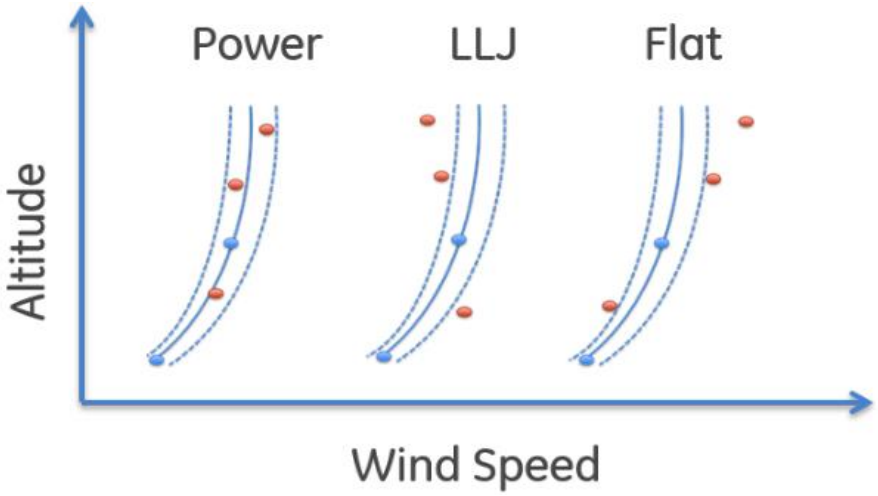


Figure 2 Wind speed profiles of Power Law, Low Level Jet (LLJ) and Flat

Analysis of field data showed that wind speed follows three main types of profiles as in Figure 2 and Table 17 (Appendix A). In Figure 2, profile on left fits well with *Power Law*.

Middle profile is Low-Level-Jet (*LLJ*), which is more curvature and smaller wind speed at the top. Right profile is *Flat* and wind speed at the top is larger. Blue dots indicate wind speeds at lower blade tip height and at hub, respectively. Red dots are speeds at other heights that are NOT available for most installed wind turbines. Blue lines are obtained using two blues dots and power law fitting.

Even though the wind shear can be critical, it is still not considered in optimizing wind turbine power output. To optimize wind turbine output, GE Wind has come up a new idea of using existing load sensor data to estimate wind shear as illustrated in Figure 1. The objective of this case study is to provide a recommendation on whether or not GE Wind should launch an NPI program to offer a "Virtual Wind Shear Sensor".

The current case study will provide detail analysis on three parts:

- **Part A:** Develop an algorithm to classify wind speed profile. The first task is to develop and evaluate a classifier to see how well a speed profile can be determined from the load sensor data. Here, we will use the load sensor data X from turbine sensors (14 parameters) to classify the wind speed profiles (Y estimator output): Power Law (0), LLJ (1), Flat (2), Other (3). The classifier accuracy will be evaluated by using the training and test data sets.

- **Part B:** Estimation of alpha from load sensor data. If the speed profile follows a power law (ShearTypeClass = 0), the shear (alpha) can be calculated by using wind speeds at altitudes of 38m and 78.7m as,

$$alpha = \ln\left(\frac{V_{38m}}{V_{78.7m}}\right) / \ln\left(\frac{38}{78.7}\right) \quad (1)$$

  The wind speed profile for Power Law can then be calculated as:

$$V(h) = V(HH) * \left(\frac{h}{HH}\right)^{\alpha} \quad (2)$$

  The task is to create a predictive model for alpha from the load sensor data and determine how well it estimates the actual alpha. Here, the same load sensor inputs (14 parameters) will be used to infer a new continuous variable Y (alpha).

- **Part C:** Estimation of shear when speed profiles do not follow a power law. For speed profiles labeled as ShearTypeClass = LLJ (1) and Flat (2), create different models to estimate wind shear.

The current project will follow the common process in data science as Obtain, Scrub, Explore, Model and Interpret (OSEMN). The data obtaining and scrubbing processes are ignored due to the provided labeled dataset from Xu and Evans (2016).

The program used to complete the case study is R in RStudio (0.99.903). Major R analysis packages used are: ggplot2, caret, caTools, glmnet, rpart, nnet, kernelab, randomForest and xgboost.

2 Exploratory Analysis and Data Visualization

2.1 Overview

To prepare modeling, the data was cleaned as following:

- converting the *datetime* to R time value
- adding the calculated wind shear(alpha) according to Eq.(1)
- adding factor "day" from *datetime* (12 factors)
- adding factor "hour" from *datetime* (24 factors)

A quick summary of the dataset is shown in Table 1 and Table 2. It contains 8046 rows and 24 columns including added columns of "alpha", "day" and "hour". The first column is *datetime,* which is the timestamp of the data. The timestamp interval is 2min. The data was continuously sampled for ~12 days. The next 5 columns from "*m38"* to "*m122"* (5 in total) are wind speed from the LIDAR sensors. The following 14 columns from "*RPM_0P"* to "*pitch_col_0P"* are load sensor data, which are main predictors for classification and regression models. The following column is the labeled *ShearTypeClass* by the field engineer. Last three columns are the added "alpha", "day" and "hour". "alpha" is the calculated wind shear from Eq.(1). Table 1 also indicated no missing data point found. Calculated skewness and kurtosis indicated that most predictors are not NORMAL distribution (skewness =0, kurtosis = 3).

Table 1 Summary of provided dataset

	vars	n	mean	sd	skew	kurtosis	se	Q0.25	Q0.5	Q0.75
datetime*	1	8046	NaN	NA	NA	NA	NA	NA	NA	NA
m38	2	8046	7.70	2.25	0.24	0.30	0.03	6.18	7.62	9.13
m58	3	8046	8.91	2.26	0.13	0.08	0.03	7.30	8.91	10.49
m78	4	8046	9.85	2.40	0.05	-0.37	0.03	8.06	9.95	11.65
m103	5	8046	10.73	2.69	-0.01	-0.66	0.03	8.63	10.87	12.72
m122	6	8046	11.26	2.92	-0.05	-0.71	0.03	8.90	11.38	13.48
RPM_0P	7	8046	13.45	1.55	-1.89	2.70	0.02	13.13	14.31	14.32
nodd_0P	8	8046	-612.93	296.03	0.10	0.72	3.30	-715.26	-691.72	-445.53
nodd_3C	9	8046	48.31	94.72	1.82	5.73	1.06	-6.37	27.96	80.82
nodd_3S	10	8046	93.96	93.75	1.15	2.71	1.05	37.97	78.90	131.31
pitch_d_0P	11	8046	0.41	0.77	0.86	0.03	0.01	0.00	0.00	0.93
pitch_q_0P	12	8046	0.11	0.24	0.57	3.45	0.00	0.00	0.00	0.25
pitch_d_3C	13	8046	-0.08	0.10	-1.35	1.64	0.00	-0.13	-0.03	0.00
pitch_d_3S	14	8046	0.00	0.05	-0.47	9.96	0.00	-0.01	0.00	0.01
yaw_0P	15	8046	58.31	84.09	0.51	1.35	0.94	13.12	27.15	111.30
yaw_3C	16	8046	57.12	80.86	1.14	2.66	0.90	7.17	42.67	91.48
yaw_3S	17	8046	-43.51	52.23	-0.51	2.37	0.58	-67.37	-38.36	-14.47
P_el	18	8046	1592.62	676.69	-0.47	-1.13	7.54	1015.37	1770.10	2283.88
V_estim	19	8046	9.58	2.24	-0.06	-0.39	0.02	7.93	9.79	11.29
pitch_col_0P	20	8046	2.59	2.88	1.42	1.64	0.03	0.24	1.74	4.02
ShearTypeClass*	21	8046	1.29	0.63	2.34	5.21	0.01	1.00	1.00	1.00
alpha	22	8046	0.36	0.24	1.65	9.11	0.00	0.18	0.34	0.51
day*	23	8046	6.52	3.25	0.00	-1.17	0.04	4.00	7.00	9.00
hour*	24	8046	12.51	6.87	0.00	-1.18	0.08	7.00	13.00	18.00

Table 2 Summary of factor predictors

datetime	ShearTypeClass	day	hour
Min. :2015-06-17 10:30:00	Power:6346	2015-06-18: 720	13 : 361
1st Qu.:2015-06-20 05:31:30	LLJ :1177	2015-06-19: 720	11 : 360
Median :2015-06-23 00:34:00	Flat : 394	2015-06-20: 720	12 : 360
Mean :2015-06-23 00:34:00	Other: 129	2015-06-21: 720	14 : 350
3rd Qu.:2015-06-25 19:36:30		2015-06-22: 720	10 : 345
Max. :2015-06-28 14:39:00		2015-06-23: 720	00 : 330
		(Other) :3726	(Other):5940

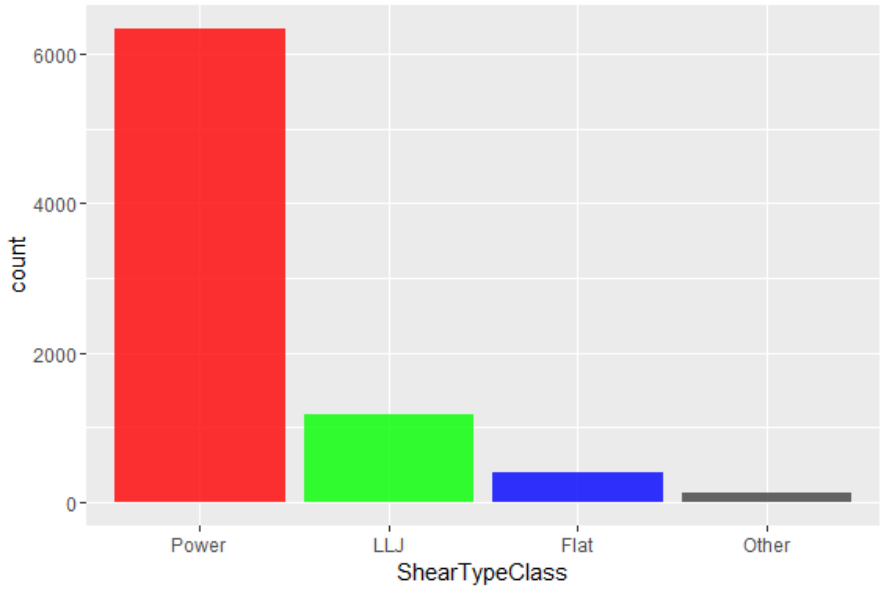
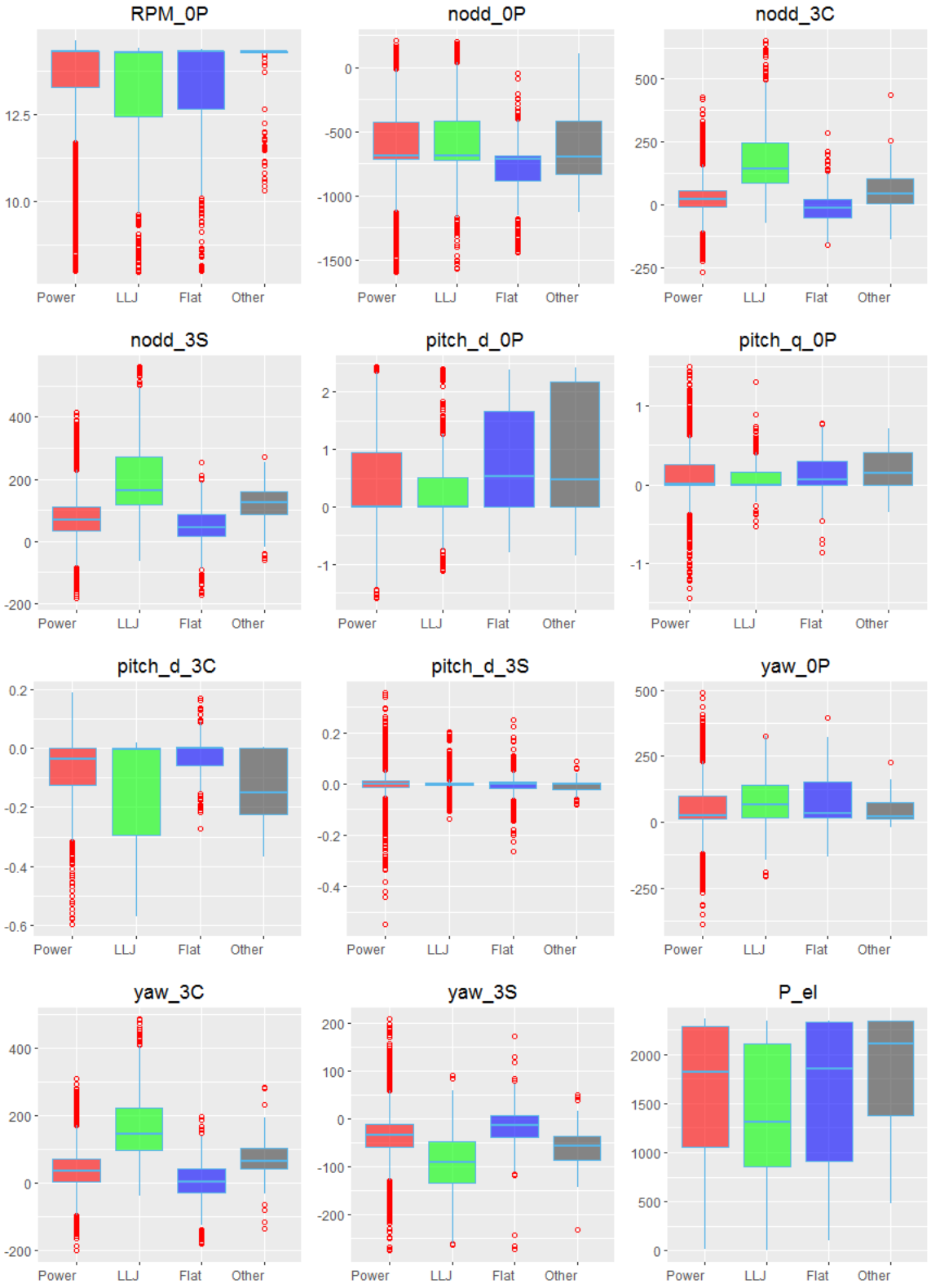


Figure 3 Wind speed profile distribution of 8046 samples.
Power (6346, 78.9%), LLJ (1177, 14.6%), Flat (394,4.95%), Other (129, 1.6%)

The labeled wind shear class distribution is shown in Figure 3. As seen in Table 1, Table 2 and Figure 3, the samples were labeled as Power (6346, 78.9%), LLJ (1177, 14.6%), Flat (394,4.95%), Other (129, 1.6%) out of 8046 total samples. The provided dataset is highly imbalanced due to the dominant Power Law (78.9%) class.

2.2 Distribution

To examine the data distribution, Figure 4 showed the boxplots for the 14 load sensor predictors and wind shear (alpha). In the boxplot, the bottom and top of the box are the lower (25%) and upper (75%) quartiles, respectively. The horizontal line near the middle of the box is the median (50%). Read dots indicate potential outlier data points. Figure 4 clearly indicated some potential outliner data in 14 predictors and alpha. The potential outlier data points were kept in the current case study since there is not enough information to filter. Figure 4 also indicated that load sensors have very different units and signal amplitudes. Also, most predictors do not have zero mean. Data pre-processing, like centering and scaling, will be needed for prediction models. It should also be noted that wind shear "alpha" was calculated for all four wind shear classes. But by definition, it only applies to the Power Law profile as shown in Figure 2.


RPM_0P
nodd_0P
nodd_3C
nodd_3S
pitch_d_0P
pitch_q_0P
pitch_d_3C
pitch_d_3S
yaw_0P
yaw_3C
yaw_3S
P_el
Power
LLJ
Flat
Other

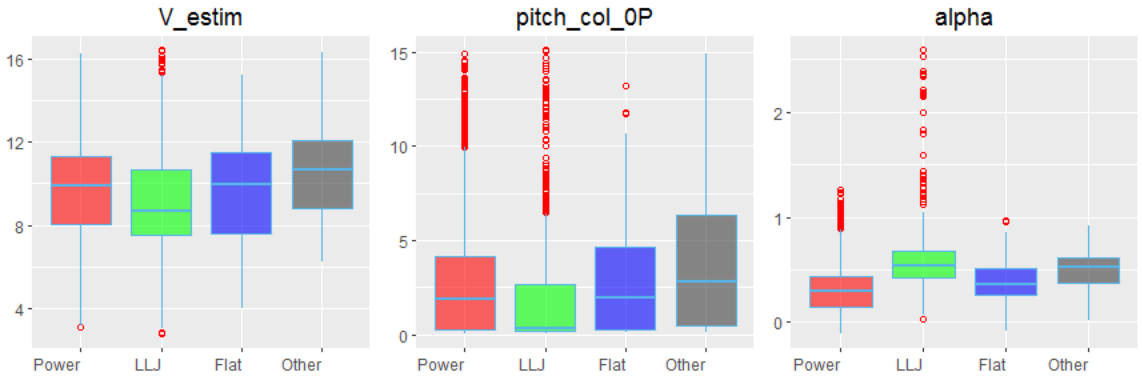
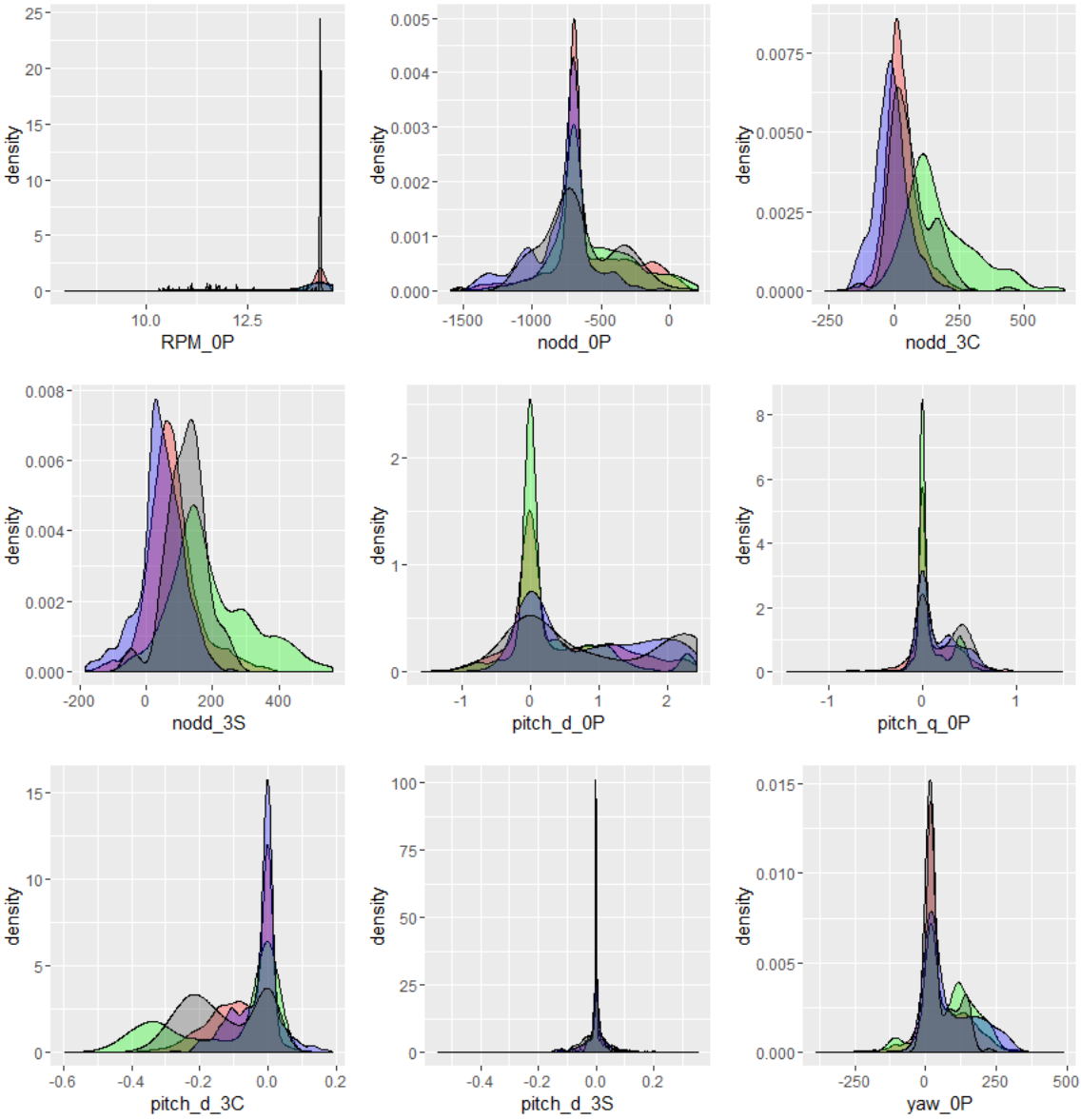


Figure 4 Boxplot for 14 wind load predictors and wind shear (alpha). The bottom and top of the box are the lower (25%) and upper (75%) quartiles. The horizontal line near the middle of the box is the median (50%). Read dots indicate potential outlier data points.



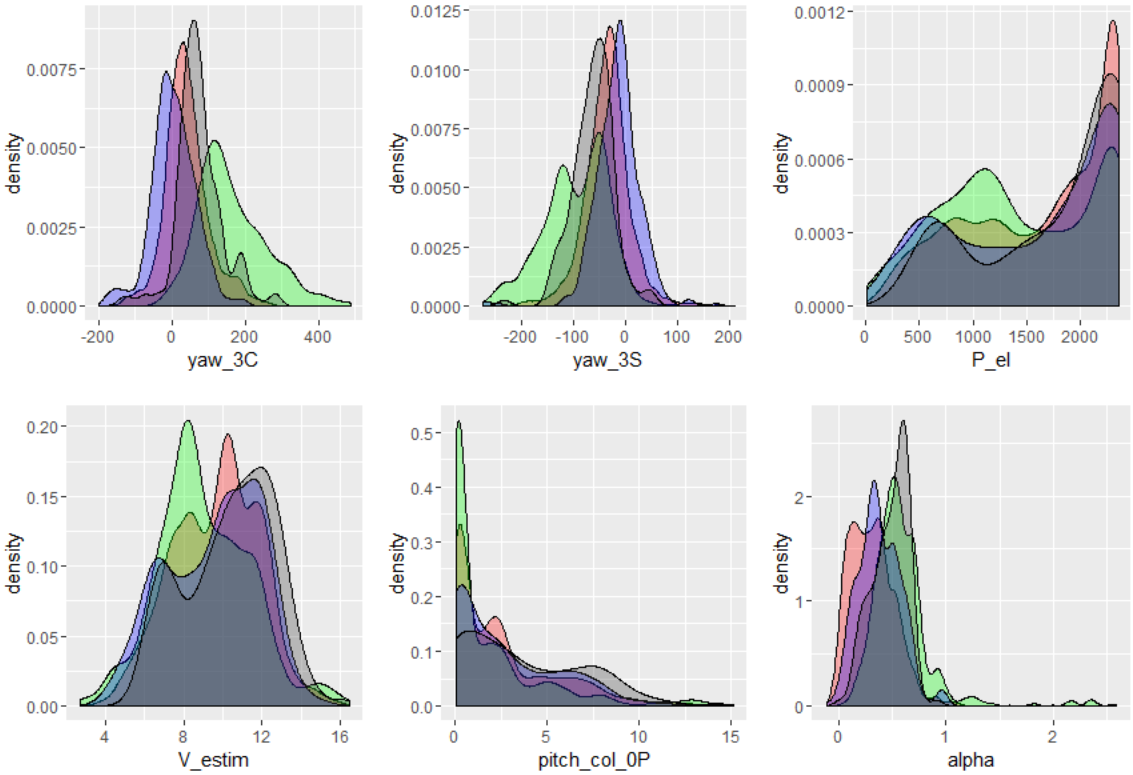


Figure 5 Distribution of 14 load sensor predictors and wind shear (alpha). Red is for wind profile of Power Law, green is for LLJ, blue is Flat and black is Other.

To further examine the data distribution, Figure 5 showed the histogram plots in different wind shear classes for the 14 load sensor predictors and the wind shear (alpha). Clearly, most all of the predictors do not follow the NORMAL distribution. Some histograms are strongly skewed, like "RPM_0P", "nod_3C" and "pitch_col_0P", which agree with skewness values in Table 1. Predictors of "RPM_0P" "nodd_0P, "pitch_d_0P", "pitch_q_0P", "pitch_d_3C", "pitch_d_3S" and "yaw_0P" all have narrow dominant values. They may correspond to a preferred condition that the wind turbine operates. Due to imbalanced data nature in wind shear classes, the distribution in wind speed profile of LLJ, Flat and Other may not be repressive (not enough data).

2.3 Time Series

As can be seen from data summary in Table 1 and Table 2, provided data was continuously sampled for ~12days at 2-min interval. It is time series in nature. Figure 6 showed the time series plots for 2 selected predictors ("nodd_0P", "pitch_d_0P") and wind shear (alpha) during the measurement time period. The data is colored by wind shear class (ShearTypeClass). Red is for wind profile of Power Law, green is for LLJ, blue is Flat and black is Other. Time series plot for other predictors are similar to Figure 6 (not shown).

Time series plots are reasonably good in general and no clearly abnormal data points. The time-series plot also clearly indicated significant differences of signal amplitude. This kind of differences may lead to biased predictors selection in model tuning. Predictor pre-processing will be necessary before model tuning.

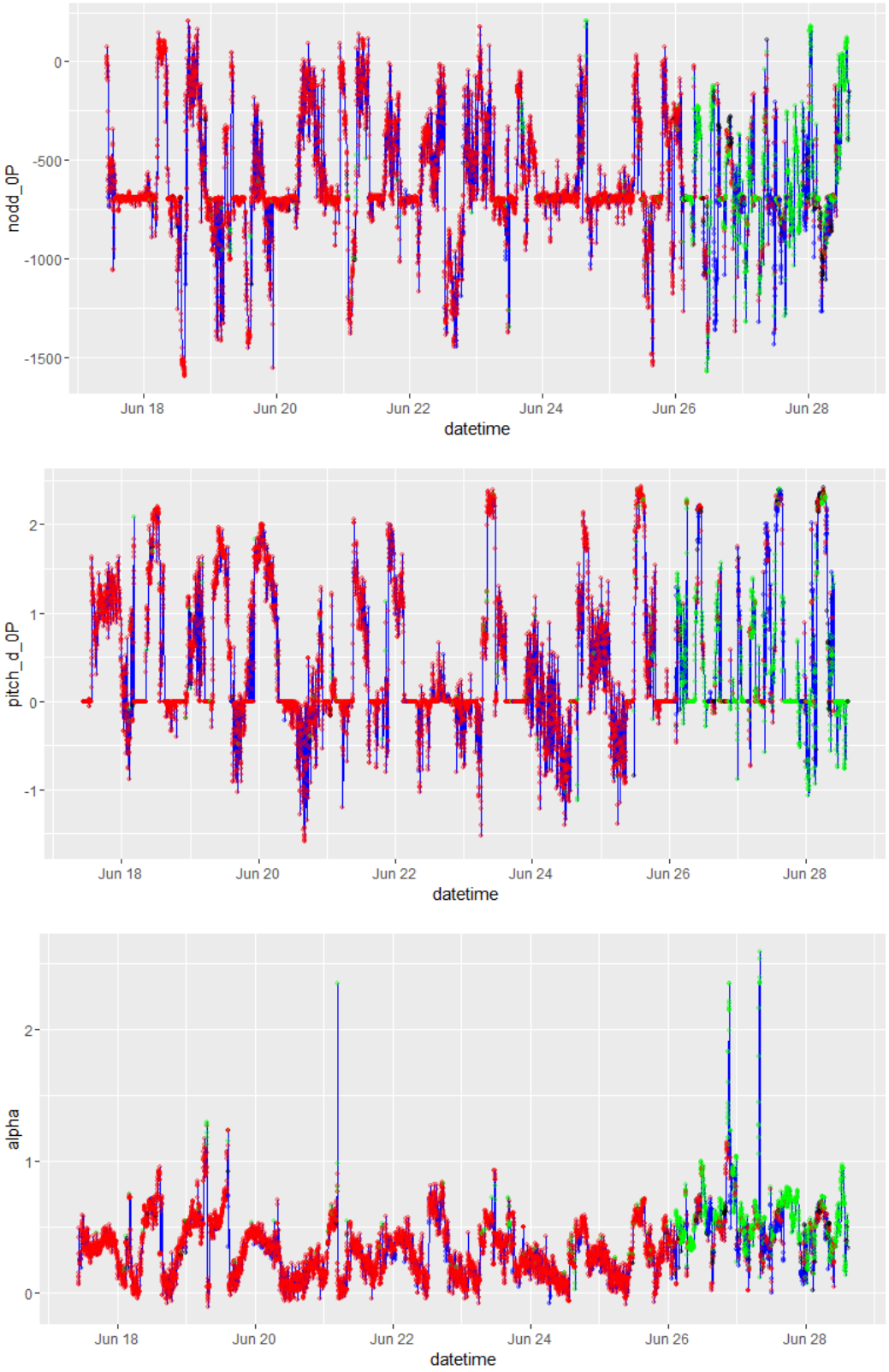


Figure 6 Time series plot of 2 typical load sensor predictors and wind shear (alpha). X-axis is the time stamp at 2-min interval. Y-axis is the predictor value.

Figure 6 also indicated that most of wind shear classes are Power Law for the first 9 days and LLJ for the last 3 days. To further explore the wind shear class distribution, Figure 7 showed the wind shear class counts in each test day using the added factor ("day"). More than 50% of data in last 3 days are of LLJ class. There is not enough information to evaluate whether this irregular wind shear class distribution is normal or not.

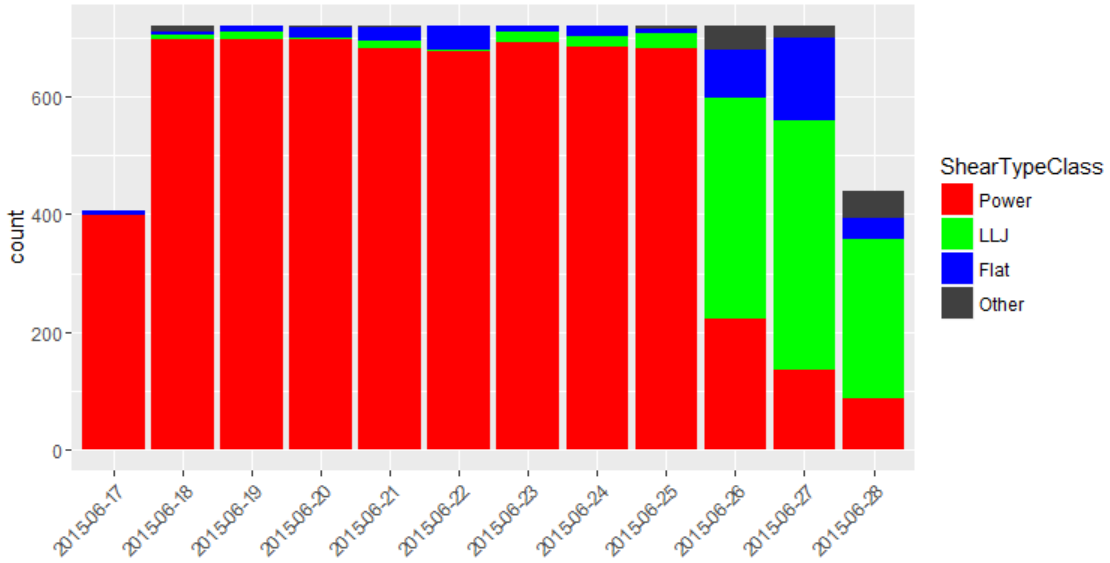


Figure 7 Distribution of wind shear class (ShearTypeCLass) at each test day.

Another inspection of test data was shown in Figure 8 for wind shear class counts in each test hour of the day using the added factor "hour". There is no clear trend of the data as function of the test hour.

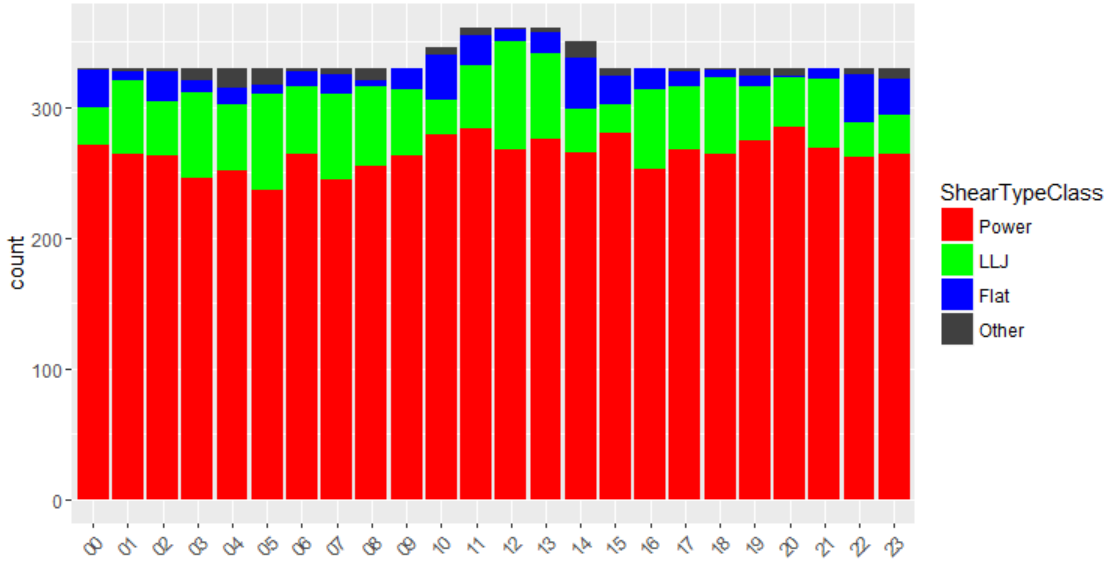


Figure 8 Distribution of wind shear class (ShearTypeClass) at each test hour of the day.

2.4 Wind Speed Profile

Wind speed sensor data is critical for wind shear estimation. Wind speed profile was labeled by field engineer as "ShearTypeClass". Wind speed profile information will be critical in Part C for wind shear estimation of minority classes of LLJ, Flat and Other.

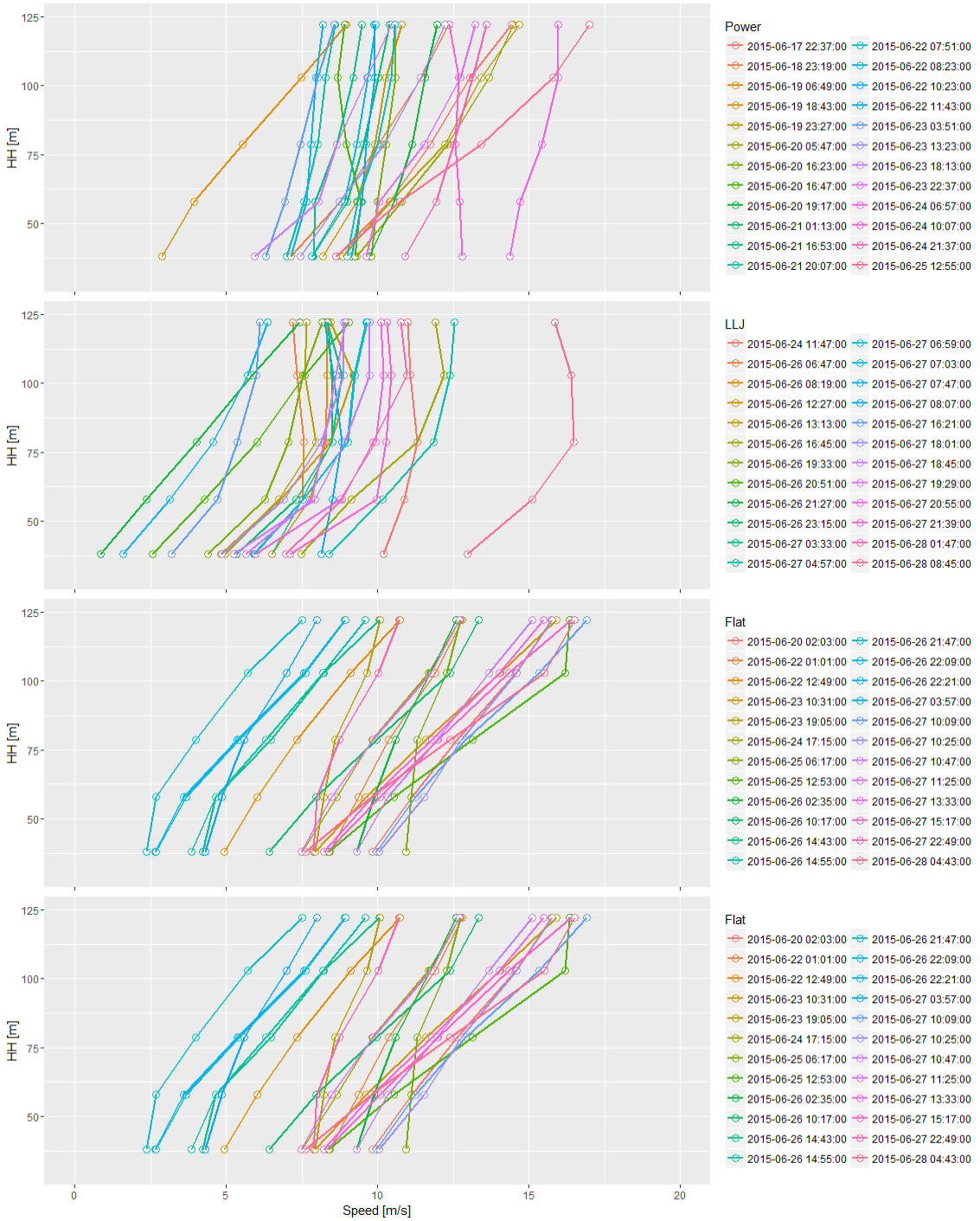


Figure 9 Illustration of 24 randomly sampled wind speed profiles for each wind shear class.

Figure 9 showed 24 randomly selected wind speed profiles for each labeled. The "Power Law" profile model was provided by the field engineer as in Eq.(1) using wind speed sensor data at hub height (78.7m) and lower tip height (38m). There are no provided wind speed profile models for LLJ, Flat and Other. However, Figure 9 clearly showed that the "Flat" speed profile can be approximated as a straight line. Figure 9 also indicated that some profiles in Power and LLJ can be potential Flat class. Improvements for ShearTypeClass labeling seem feasible.

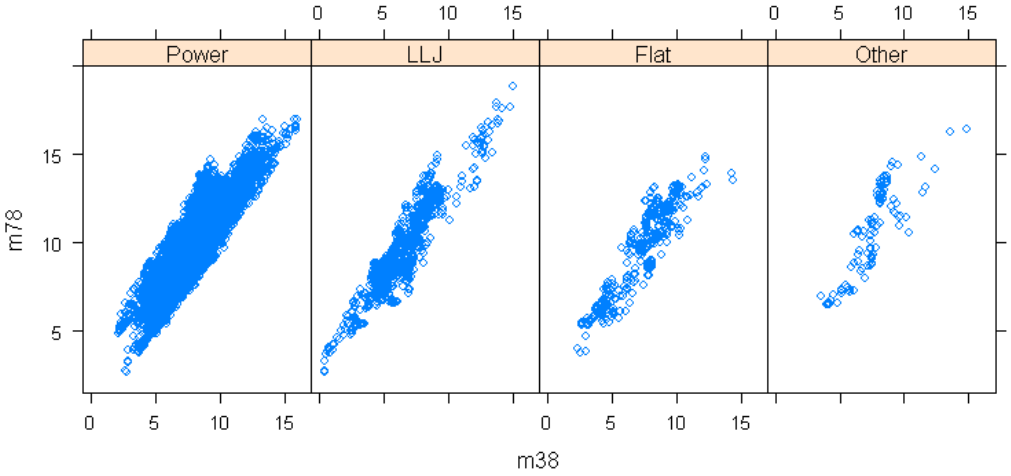


Figure 10 Scatter plot of wind speed sensor data "m38" and "m78" at each wind profile class.

Since most installed wind turbine only have speed sensors at hub height (78.7m) and lower tip height (38m). Therefore, it is customary in the industry to use these two speeds to estimate wind speed profile. Figure 10 showed the scatter plots of the wind speed senor data of "m38" and "m78" for all wind shear classes and strong linear correlations are apparent.

2.5 Correlation

Correlation of the predictors can be important for modeling. To examine the correlation effects, the scatter plots and correlation values of 14 load sensor predictors and the calculated wind shear (alpha) were shown in Figure 11.

Lower left of Figure 11 showed the scatter plots of every 2-predictor combinations from 14 predictors and the calculated wind shear (alpha). Upper right of Figure 11 showed the absolute correlation coefficients for the 2-predictor pairs. The font size is proportional to the absolute correlation coefficient. The diagonal of Figure 11 showed the histogram plots of 14 predictors and alpha. The distributions of some predictors, "P_el" for example, are very skewed and far from NORMAL distribution.

Figure 11 also clearly showed strong correlations among some load sensor predictors. For example, the correlations between ("RPM_0P", "P_el"), ("RPM_0P", "V_estim"), ("nod_3S, yaw_3C"), ("P_el", "V_setim"), ("P_el", "pitch_col_0P") and ("V_estim", "pitch_col_0P") are all higher than 0.75. They were highlighted red in Figure 11. The max correlation value for wind shear (alpha) is 0.51 with signal "nodd_0P". Some of high correlation predictors will be removed for modeling, which will be discussed in Section 3.

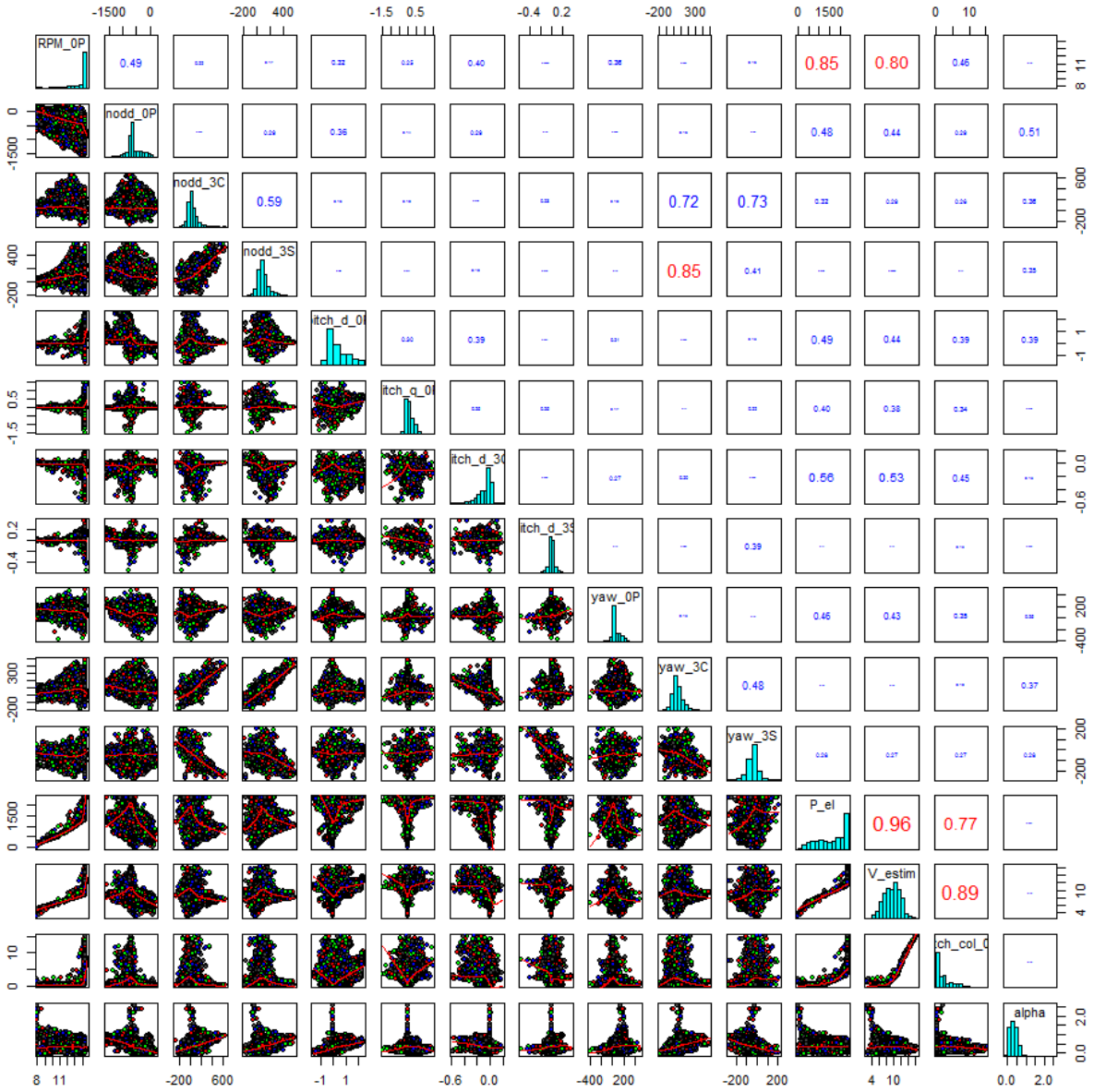


Figure 11 Correlation plot of 14 load sensor predictors and wind shear (alpha). Lower left shows scatter plots of every 2-predicotr combinations. Upper right shows absolute correlation coefficients. Diagonals are histogram plots.
Red is Power Law, green is LLJ, blue is Flat and black is Other.

The correlations among wind speed sensor data can affect wind speed profile modeling, especially in Part C. Figure 12 showed that there are very strong correlations among the 5 wind speed sensors. The absolute correlation coefficients are close to 1, especially for adjacent wind speed sensors. Wind speed sensor information of "m38" and "m78" were used to estimate the wind shear (alpha) in Eq.(1). The correlation for the wind speed sensor "m38" and "m78" is 0.86.

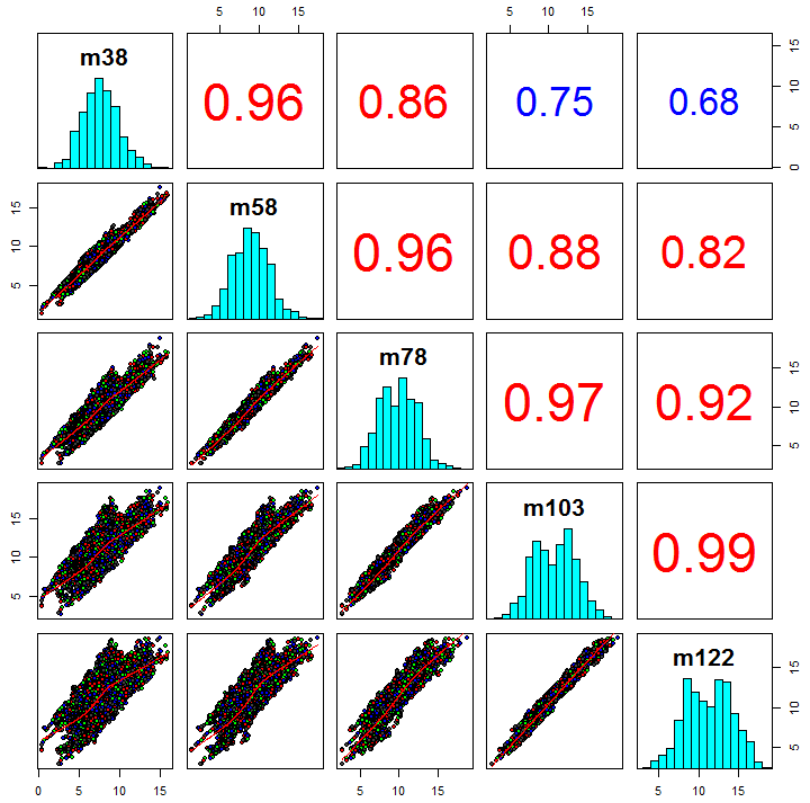


Figure 12 Correlation of wind speed sensor data. Lower left shows the scatter plots of every 2-sensor combinations. Upper right are absolute correlation coefficients. Diagonals are histogram plots.

2.5 Data Summary

The provided dataset is of reasonably good quality with no missing data point (Table 1) and no clear abnormal trend (Figure 6). However, there are some potential issues of the current dataset and here is a brief summary,

- The data is highly imbalanced. Power Law class data is dominant (78.9%) of the overall data. (Table 1 and Figure 3)
- There are potentially strong outlier data in 14 load sensor predictors and wind shear (alpha). (Figure 4 and Figure 5)
- Time series plots indicated clear biased wind shear class distribution in field test days. Most data Power Law class before 2015-06-26, and LLJ for test days after 2015-06-26. (Figure 6 and Figure 7)
- There are strong correlations among 14 load sensor predictors. Correlations between ("RPM_0P", "P_el"), ("RPM_0P", "V_estim"), ("nodd_3S", "yaw_3C"), ("P_el", "V_setim"), ("P_el", "pitch_col_0P") and ("V_estim", "pitch_col_0P") are higher than 0.75. (Figure 11)

- Correlations among wind speed sensors are close to 1, especially for adjacent sensors. (Figure 12)
- Signal units and magnitudes of 14 wind load sensors are significantly different. (Table 1)
- The provided ShearTypeClass labeling can be improved by inspecting wind speed profiles for each shear class (Figure 9).

3 Modeling

The modeling process flow chart for current case study was shown in Figure 13. The feature engineering was the first step to remove high correlation predictors. The filtered dataset was then split into training and test sets. Training data was used to tune model hyperparameters based on certain metrics. Predictions were then made by models with optimized hyperparameters using test data. Model performances were evaluated by comparing differences between predictions and test labels. Final model was down-selected with best ranking with specific selection metric.

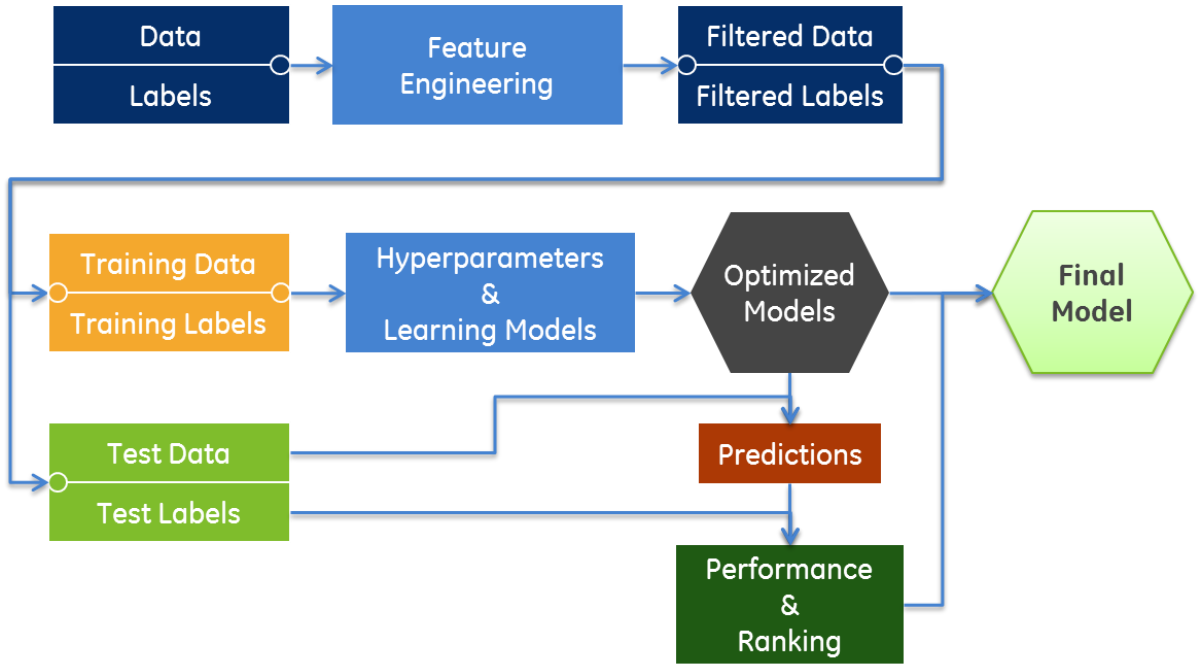


Figure 13 Model selection process

3.1 Data Preparing

3.1.1 Data Filtering

Three predictors, **"yaw_3C", "P_el" and "V_estim**", were removed based on the correlation threshold of 0.75. Correlations of filtered dataset with 11 predictors were shown in Figure 14 with max absolute correlation coefficient of 0.73 between "nodd_3C" and "yaw_3S". Here

the correlation is represented by a circle and its size is proportional to the absolute correlation coefficient. Negative and positive correlations are marked as red and blue respectively. These 11 predictors will be used in modeling.

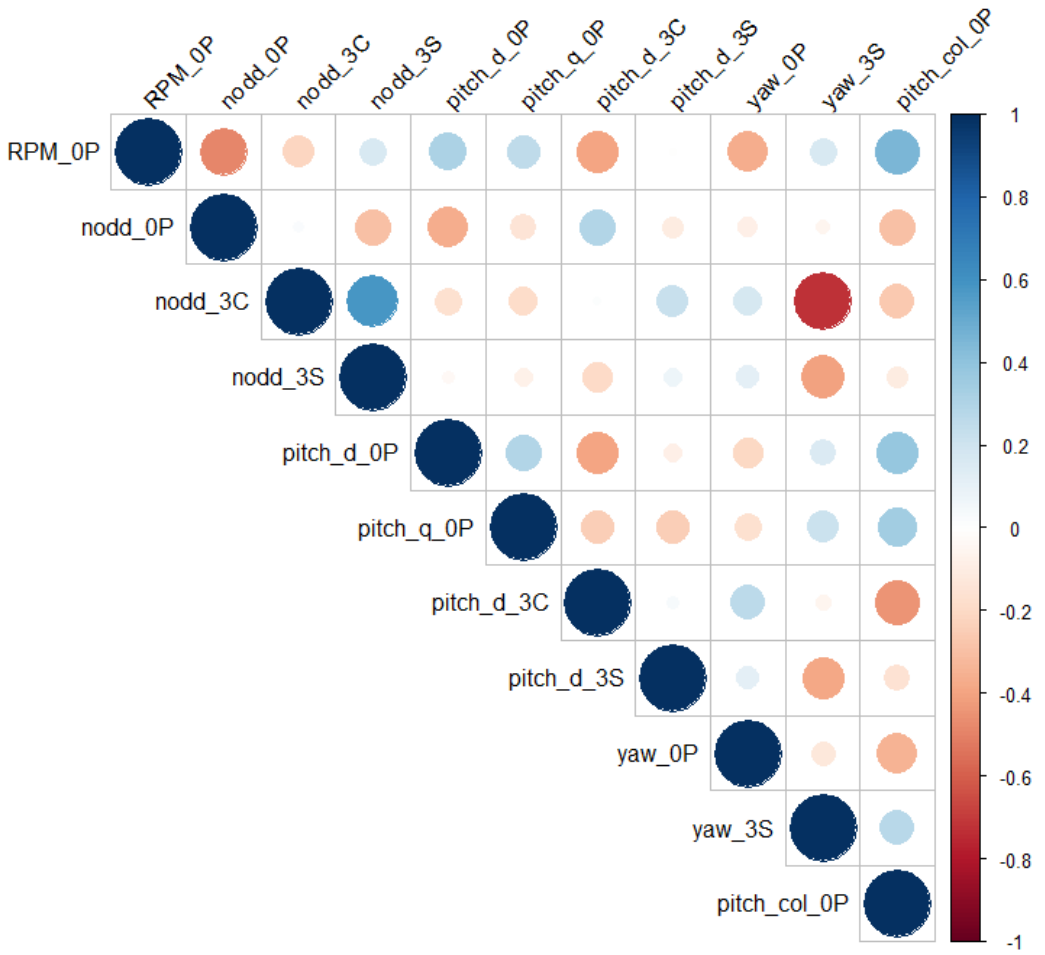


Figure 14 Correlation of filtered load sensor data (11 predictors). Blue/Red indicate positive/negative correlations. Absolute coefficients are proportional to circle size.

3.1.2 Data Splitting

Data split for training and test is 75% and 25%, respectively. Training data was used multiple times for model hyperparameters optimization by repeated CV process. Test data was used only once for model evaluation.

In Part A, the split of classification data set was shown in Figure 15. It can be seen that roughly the same percentage of data is in test dataset across the wind shear classes. This is critical for model training and evaluation.

In Part B, the split of wind shear (alpha) for wind speed profile of Power Law is shown in Figure 16. Splits for training and test data for wind speed profiles of LLJ and Flat in Part C are similar to Figure 16 (not shown).

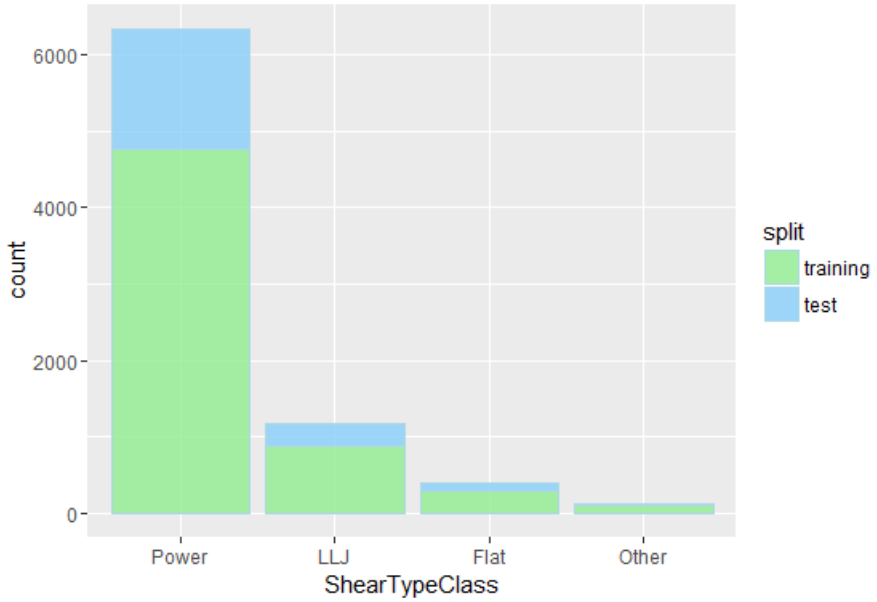


Figure 15 Split of training and test data for ShearTypeCLass (Part A).

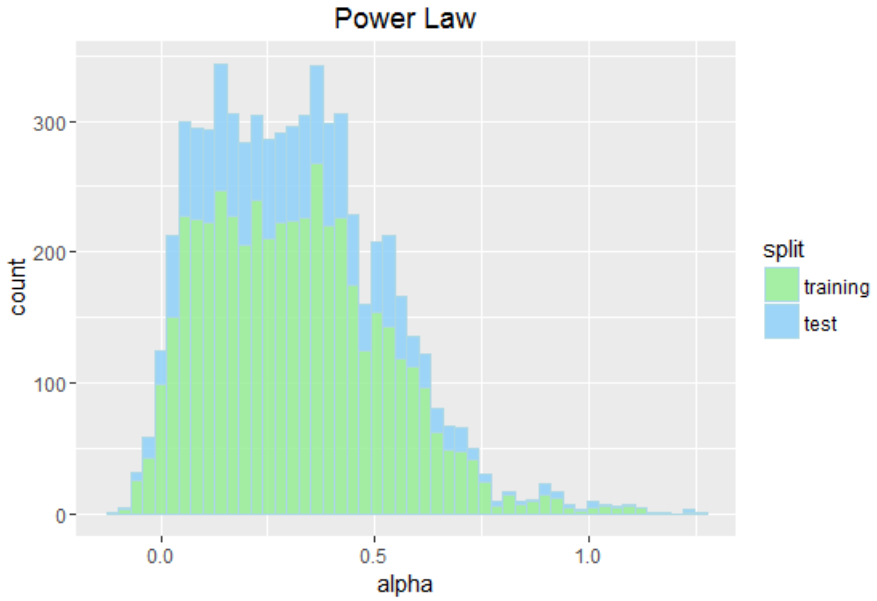


Figure 16 Split of training and test data for model parameter alpha with Power Law (Part B).

3.1.3 Data Preprocessing

All load sensor predictors will be "centered" and "scaled" before tuning classification and regression models to remove impacts of data units and amplitudes. This will avoid potential bias predictor selection in model tuning process. It should be noticed that preprocessing is not necessary for tree based algorithms.

3.1.4 Cross Validation

5-repeat of 10-fold Cross Validation (CV) was adopted to optimize model hyperparameters. The combined repeated CV process leads to 50 runs for each hyperparameter combination. For consistent model comparison, training and CV samples were fixed for all classification models in Part A. This process was also applied to regression models in Part B and C.

3.1.5 Imbalanced Classes

Imbalanced data typically refers to a classification problem where the classes are not represented equally, which is an issue in the current case study. Several common approaches to deal with imbalanced data are (Kuhn 2016): collecting more data, using different performance metrics, resampling dataset, synthetic samples, different algorithms and cost weighted models. In the current study, different performance metrics and different algorithms were first studied in section 4. After down-selecting the best classifier, synthetic sample techniques were then applied to check the impact of data imbalance on model performance. Two synthetic algorithms were explored: Synthetic Minority Over-sampling TEchnique (SMOTE, Chawla et al. 2002) and Randomly Over Sampling Examples (ROSE). SMOTE and ROSE down-sample the majority class and synthesize new data points in the minority classes. The original R packages (DMwR and ROSE) mainly work for binary classification and were modified for multi-classification modeling. It should also be noted that synthetic sample techniques were applied to training dataset only. Same test data as in Figure 15 will be used for model performance evaluation.

3.2 Modeling

3.2.1 Model Options

Supervised learning algorithms were considered in this case study since the provided data has been labeled and wind shear (alpha) was calculated from Eq.(1). Commonly used machine learning algorithms were discussed by Kuhn and Johnson (2013) and Fernandez-Delgado et al. (2014). In the current case study, the selected models are summarized in Table 3. These algorithms are capable for both classification and regression problems

Table 3 Summary of supervised learning models in the case study

Model Name	Method	R Package	Hyperparameters
Generalized Linear Model (GLM)	glmnet	glmnet	alpha, lambda
Classification and Regression Tree (CART)	rpart	rpart	cp
Neural Network (NN)	nnet	nnet	size, decay
Support Vector Machine (SVM)	svmRadial	kernlab	sigma, C
Random Forest (RF)	rf	randomForest	mtry
eXtreme Gradient Boosting (XGB)	xgbtree	xgboost	nrounds, eta, max_depth

3.2.2 Model Training

Grid search was used in this case study to find optimized hyperparameters and achieve best prediction. The hyperparameters tuning process were evaluated based on 5-repeat of 10-fold cross validation. The detail of model training using R package "caret" can be found in Kuhn (2008), Kuhn and Johnson (2013) and Kuhn(2016).

In Part A, all classification models were trained with the same training set. Classification model training metric is "Kappa" due to imbalanced predictors in wind shear classes. The overall "Accuracy" metric was also checked.

In Part B, all regression models were trained on the same training set for wind shear class of Power Law. This training set is different from that in Part A. Regression model training metric is Root Mean Square Error (RMSE). The R^2 metric was also checked.

In Part C, "RMSE" metric was used for regression models and "R^2" metric was checked.

3.2.3 Model Selection

In Part A, variation of Kappa and overall Accuracy from 5-repeat 10-fold CV process were compared based on training data. Confusion matrix was then evaluated for all classification models based on the same test dataset. The test dataset has same imbalance as original data. Kappa and overall Accuracy were calculated based on confusion matrix. F1 score for each class was also compared. Classification model with highest Kappa value was down-selected as final classifier in Part A.

In Part B, variation of RMSE and R^2 from 5-repeat 10-fold CV process using training data were compared first. RMSE and R^2 metrics were then evaluated for all regression models using same test dataset. Regression model with lowest RMSE value was down-selected as the final model in Part B.

In Part C, only the best and worst regression models from Part B were considered. RMSE and R^2 metrics calculated from test dataset was used to evaluate model performance.

4 Interpreting Results

4.1 Part A - Wind Speed Profile Classification

4.1.1 Baseline Accuracy

The most common outcome in the provided dataset is for wind shear class of Power Law (ShearTypeClass=0). The baseline accuracy, predicting all data as Power Law class, is **0.789**.

4.1.2 Hyperparameters

Optimized hyperparameters for classification models from 5-repeat of 10-fold CV process were shown in Table 4.

Table 4 Optimized hyperparameters for Classification models (Part A)

Model Name	Hyperparameters
Generalized Linear Model (GLM)	alpha = 1 (Lasso), lambda = 3.793e-4
Classification and Regression Tree (CART)	cp = 6.905e-4
Neural Network (NN)	size = 11, decay = 0
Support Vector Machine (SVM)	sigma = 0.152, C = 8
Random Forest (RF)	mtry = 6
eXtreme Gradient boosting (XGB)	nrounds = 500, max_depth = 10, eta = 0.1, gamma = 0

Detail hyperparameter optimizations using grid search were shown in Figure 37 for Kappa metric (Appendix B.). Model with higher Kappa value is better. The optimization process for overall Accuracy is similar to Figure 37 (not shown here). Figure 37 is also important to verify that optimized hyperparameters are within the searching grid.

4.1.3 Model Performance

CART and Neural Network classification models can be visualized, as shown in Figure 17 and Figure 18.

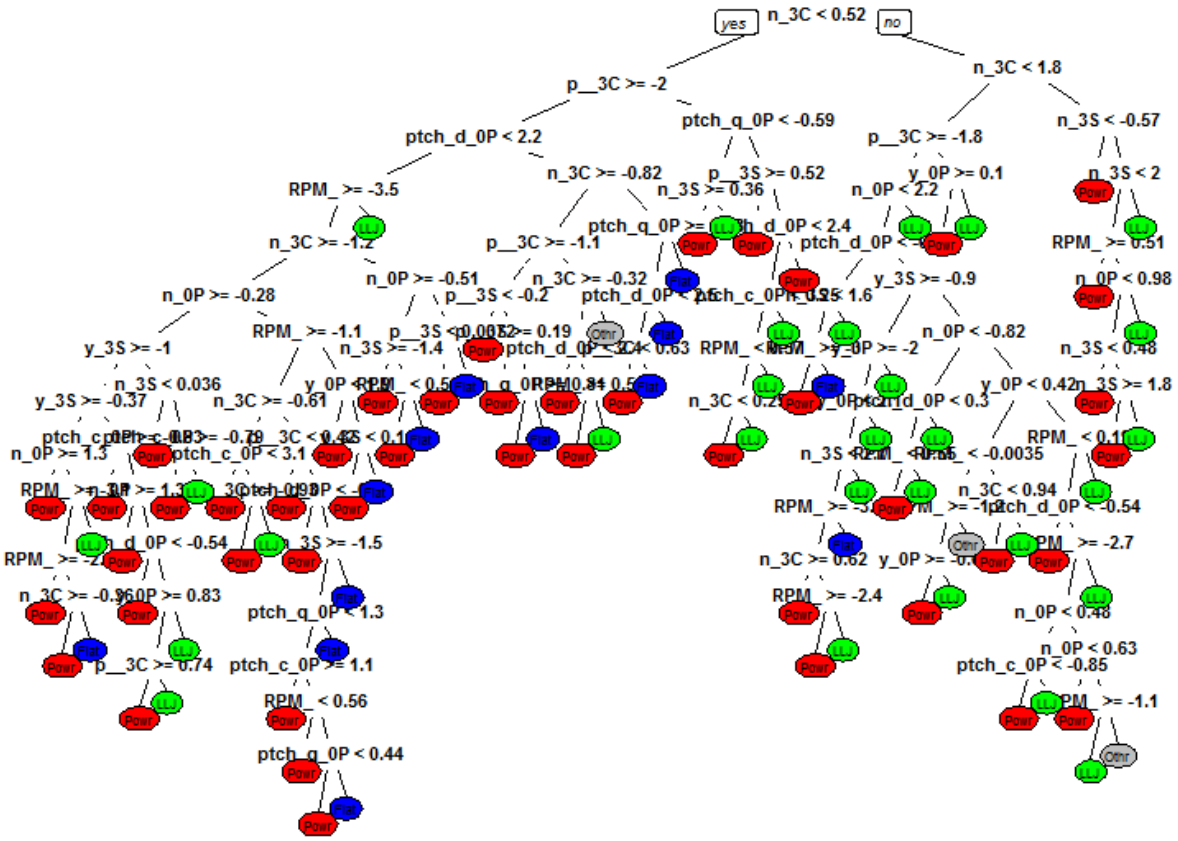


Figure 17 CART classification model (Part A)

Figure 17 showed CART tree structure with optimized hyperparameter. The root predictor is "nodd_3C". "Pitch_d_3C" and "nod_3C" are the second level split predictors. Figure 18 showed optimized Neural Network model with 1 hidden layer (11 units) and 4 class outputs.

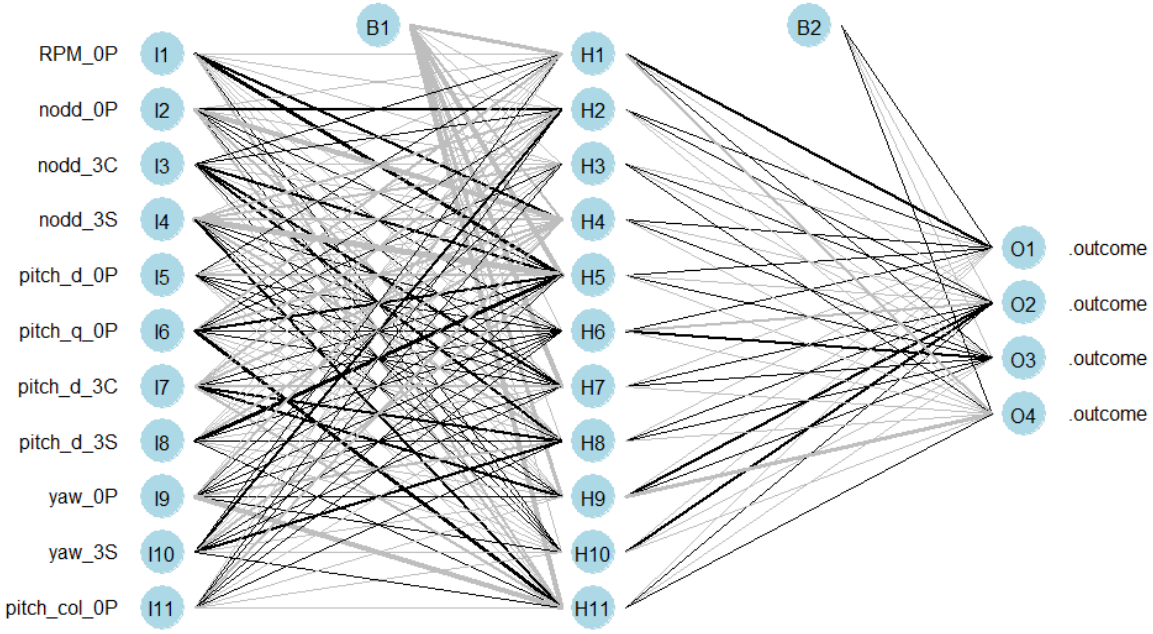


Figure 18 Neural Network classification model with 4 class outputs (Part A)

Figure 19 showed predictor importance plot for Random Forest classification model. Here, the x-axis is predictor name and y-axis is predictor's impact on the model metric. The top 3 important predictors are "nodd_3C", "pitch_d_3C" and "nodd_0P". However, there is no apparent dominant predictor in this model.

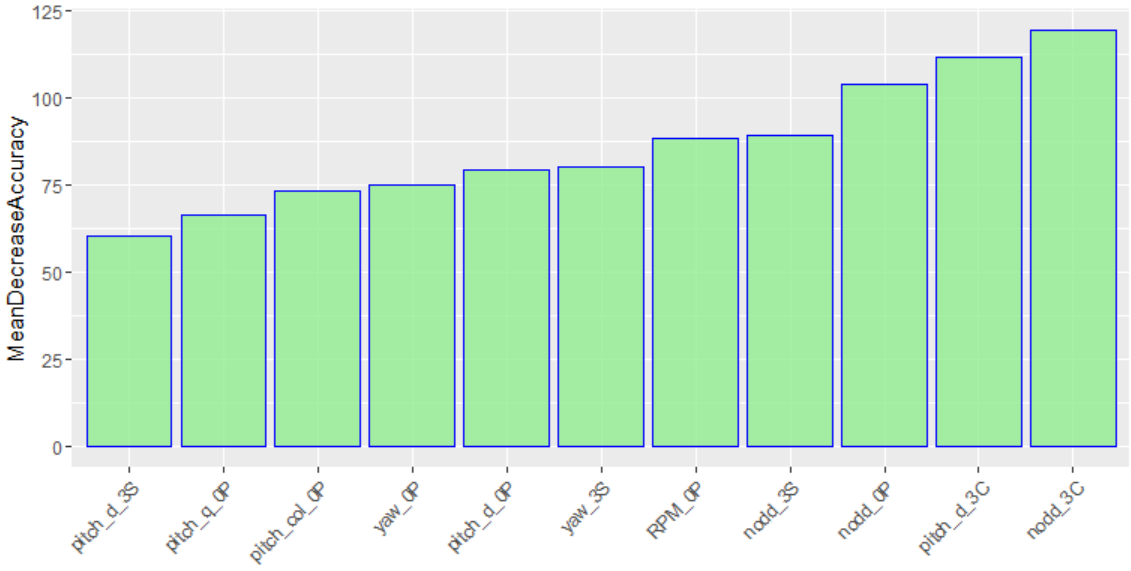


Figure 19 Variable importance for Random Forest classification model (Part A)

To evaluate model prediction performance, one approach is to check the Kappa and overall Accuracy metric variations from 5-repeat of 10-fold CV process as in Figure 20. It clearly showed metric variations from training process of 50 runs. For comparison, the baseline accuracy of 0.789 was also added in Accuracy plot. All selected classification models have better overall accuracy than baseline. XGB and GLM are top two classifiers, and CART and GLM are least performed classifiers.

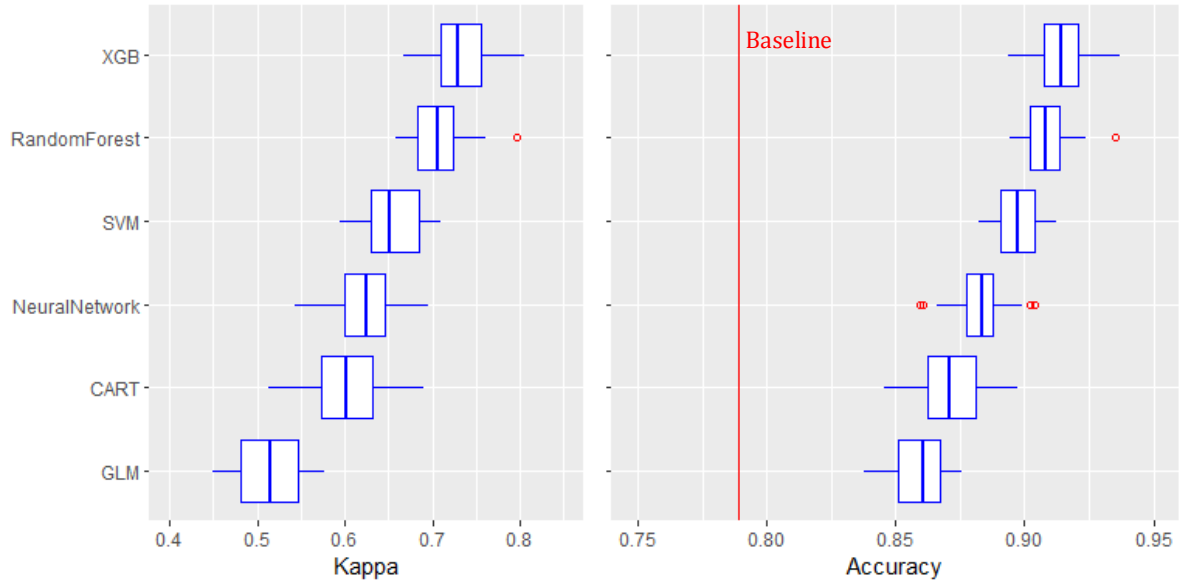


Figure 20 Classification model comparison with Kappa and overall Accuracy metrics (Part A). Model with higher Kappa and overall Accuracy is better

Table 5 Confusion Matrix for classification models (Part A).

GLM

Prediction	Reference: Power	LLJ	Flat	Other
Power	1538	124	90	29
LLJ	40	170	0	3
Flat	8	0	8	0
Other	0	0	0	0

CART

Prediction	Reference: Power	LLJ	Flat	Other
Power	1516	74	65	27
LLJ	46	217	6	2
Flat	16	1	27	0
Other	8	2	0	3

Neural Network

Prediction	Reference: Power	LLJ	Flat	Other
Power	1537	85	74	28
LLJ	38	209	2	4
Flat	10	0	22	0
Other	1	0	0	0

SVM

Prediction	Reference: Power	LLJ	Flat	Other
Power	1568	73	87	30
LLJ	15	221	2	1
Flat	1	0	9	0
Other	2	0	0	1

Random Forest

Prediction	Reference: Power	LLJ	Flat	Other
Power	1566	56	69	29
LLJ	16	238	1	1
Flat	2	0	28	0
Other	2	0	0	2

XGB

Prediction	Reference: Power	LLJ	Flat	Other
Power	1565	51	61	27
LLJ	14	241	0	2
Flat	4	0	37	0
Other	3	2	0	3

Confusion Matrix (CM) compared the predicted and true ShearTypeClass using test data, as shown in Table 5. The diagonal of the CM indicates correctly predicted classes and off-diagonal values are for incorrect predictions.

Classification model performance were summarized in Table 6 with metrics of Kappa, overall Accuracy, and F1 scores for each class. These performance metrics were calculated based on the confusion matrix in Table 5 using test data. The F1 score in Table 5 showed that GLM and Neural Network model cannot detect "Other" shear class at all. XGB has the best F1 score for all classes. CART and Neural Network have similar values of Kappa, Accuracy and F1 scores. GLM model is the worst for all metrics.

Table 6 Summary of classification model performance with test data (Part A). Performance metrics include Kappa, overall Accuracy, and F1 score for each class.

Metric	GLM	CART	Neural Network	SVM	Random Forest	XGB
Kappa	0.487	0.615	0.604	0.641	0.713	0.738
Accuracy	0.854	0.877	0.880	0.895	0.912	0.918
F1: Power	0.914	0.928	0.929	0.938	0.947	0.951
F1: LLJ	0.671	0.768	0.764	0.829	0.865	0.875
F1: Flat	0.140	0.380	0.338	0.167	0.438	0.532
F1: Other	NA	0.133	NA	0.057	0.111	0.150

It should be noted that the model section highly depends on the specific selection metric. In this case study, Kappa metric was used and the classifier performance ranking is:

```
XGB > Random Forest > SVM > CART > Neural Network > GLM
```

4.1.4 Impact of Balanced Data

As discussed in section 2.1, the provided dataset is highly imbalanced and Power Law class dominates. The confusion matrix results in Table 5 also clearly indicated the poor performance of classifier on minority classes. Therefore, it is worth to evaluate classifier's performance using balanced dataset.

Several things should be noticed for using balanced dataset to evaluate model performance.

- Only the training dataset was balanced using SMOTE and ROSE techniques as illustrated in Figure 21 and Figure 22.
- Test dataset was NOT balanced and it still has the same imbalance as original data.
- Hyperparameters for XGB model were re-optimized based on balanced training datasets from SMOTE and ROSE, respectively. These will be named as XGB-SMOTE and XGB-ROSE.

Table 7 compared original and balanced sample counts. The balanced sample count was ~600 which led to down-sampling of majority classes (Power and LLJ) and adding

synthesized new data points for minority classes (Flat and Other). The balanced dataset will be used to check down-selected XGB classifier.

Table 7 Original and balanced sample counts for classification (Part A)

Class:	Power	LLJ	Flat	Other
Training Original:	4760	883	296	97
Test Original:	1586	294	98	32
Training SMOTE:	582	582	582	582
Training ROSE:	590	590	590	590

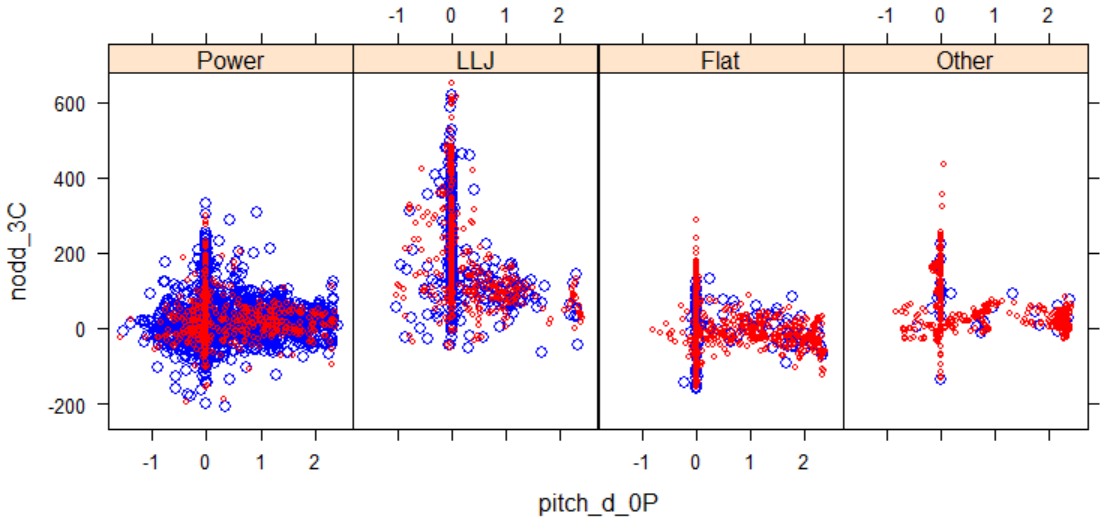


Figure 21 Comparing original and synthesized training dataset (Part A). Blue is for original training dataset and red is for synthesized data using SMOTE.

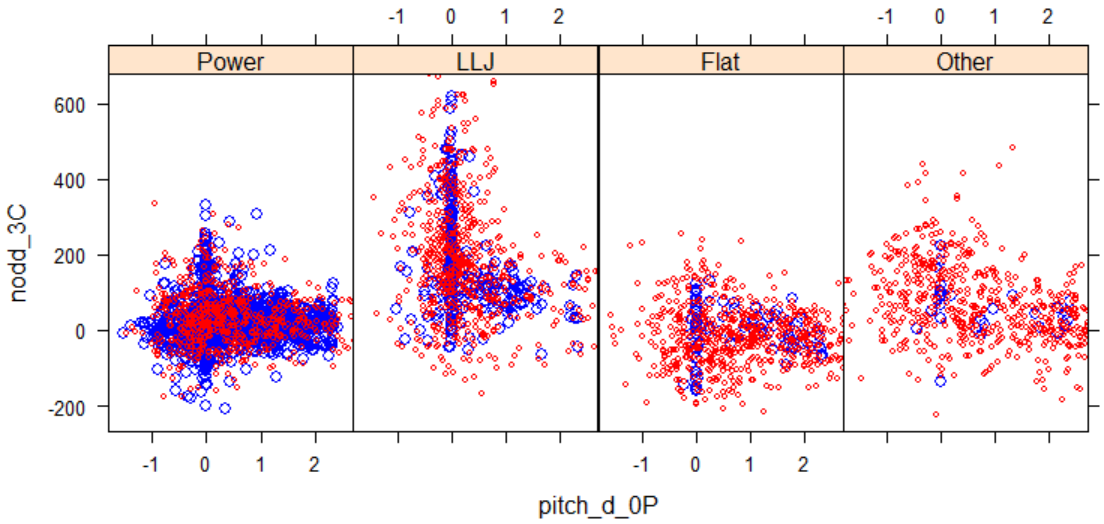


Figure 22 Comparing original and synthesized training dataset (Part A). Blue is for original training dataset and red is for synthesized data using ROSE.

Figure 21 and Figure 22 compared original training data (blue) and synthesized training data (red) from SMOTE and ROSE respectively. These figures clearly showed that SMOTE technique can create samples closely representing original data than ROSE.

Table 8 compared confusion matrix, sensitivity, specificity and F1 score for XGB model using original, SMOTE and ROSE training data. The Confusion Matrix clearly indicated that XGB model using balanced training dataset correctly captured more minority classes, especially for SMOTE technique. XGB-SMOTE model has the same sensitivity for LLJ class as the original XGB model, and much better sensitivity of minority classes of Flat and Other. Results in Table 8 implied that minority class prediction performance can be improved with balanced training dataset, at the expense of lower prediction performance of dominant class (Power).

Table 8 Impact of balanced training dataset on classifiers using SMOTE/ROSE techniques (Part A).

Model	Confusion Matrix				Sensitivity	Specificity	F1
XGB	Prediction	Power	LLJ	Flat	Other			
	Power	1565	51	61	27	0.9867591	0.6721698	0.9513678
	LLJ	14	241	0	2	0.8197279	0.9906760	0.8747731
	Flat	4	0	37	0	0.3775510	0.9979079	0.5323741
	Other	3	2	0	3	0.0937500	0.9974722	0.1500000
XGB-SMOTE	Prediction	Power	LLJ	Flat	Other			
	Power	1153	27	19	12	0.7269861	0.8632075	0.8244548
	LLJ	105	243	2	1	0.8265306	0.9370629	0.7534884
	Flat	240	6	76	3	0.7755102	0.8697699	0.3593381
	Other	88	18	1	16	0.5000000	0.9459050	0.2064516
XGB-ROSE	Prediction	Power	LLJ	Flat	Other			
	Power	1009	38	21	11	0.6361917	0.8349057	0.7572233
	LLJ	115	200	1	0	0.6802721	0.9324009	0.6557377
	Flat	280	12	70	2	0.7142857	0.8462343	0.3030303
	Other	182	44	6	19	0.5937500	0.8827098	0.1342756

Table 9 compared metrics of Kappa and overall Accuracy using original imbalanced test data. Due to the high imbalance, Kappa and overall Accuracy metrics of XGB-SMOTE and XGB-ROSE are lower than original XGB model.

Table 9 Summary of classifier performance with balanced training data (Part A).

Metric	XGB	XGB-SMOTE	XGB-ROSE
Kappa	0.738	0.470	0.347
Accuracy	0.918	0.740	0.646

Results in Table 8 and Table 9 further highlighted the importance of model evaluating metrics. If minority classes are critical, balanced training data should be used.

4.1.5 Impact of Filtered Predictors

As discussed in Section 3.1, three predictors were filtered out due to high correlation. Here, the impacts of filtered predictor on model performance was verified. In Table 10, "GLM-all" and "XGB-all" classifiers used all 14 predictors and GLM and XGB used only 11 filtered predictors. The hyperparameters of these 2 new models were re-optimized based on Kappa metric. Results in Table 10 indicated no impact on model performance due to filtered out predictors.

Table 10 Impacts of filtered predictors on classifier performance (Part A).

Metric	GLM	GLM-all	XGB	XGB-all
Kappa	0.487	0.504	0.738	0.731
Accuracy	0.854	0.856	0.918	0.916
F1: Power	0.914	0.915	0.951	0.950
F1: LLJ	0.671	0.686	0.875	0.870
F1: Flat	0.140	0.171	0.532	0.539
F1: Other	NA	NA	0.150	0.146

4.2 Part B - Estimation of alpha for Power Law Profile

4.2.1 Standard Liner Regression Models

For regression models in Part B, standard linear regression model was explored first with coefficients shown in Table 19 (Appendix C). The linear regression results indicated that predictor "pitch_d_3C" is not important in the linear regression model. Linear regression with best sub-model selection was then studied with the focus on predictor importance as shown in Figure 23 and Figure 24.

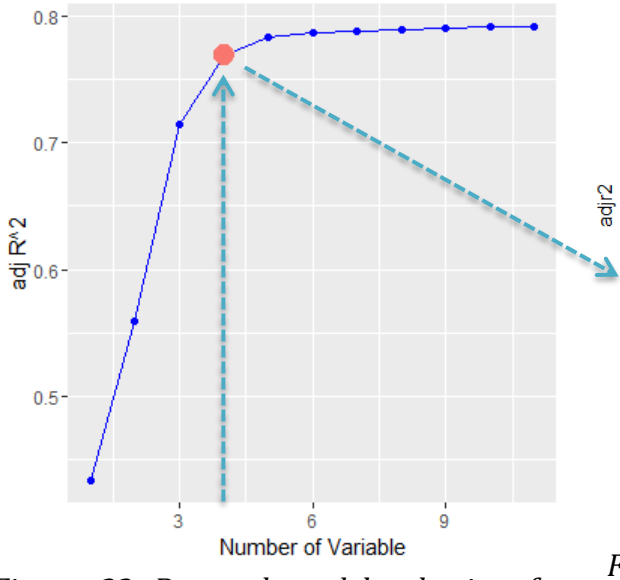


Figure 23 Best sub-model selection for linear regression. x-axis is the number of predictors used in sub-model. Y-axis is the highest adjusted R^2 value for sub-model with same number of predictors

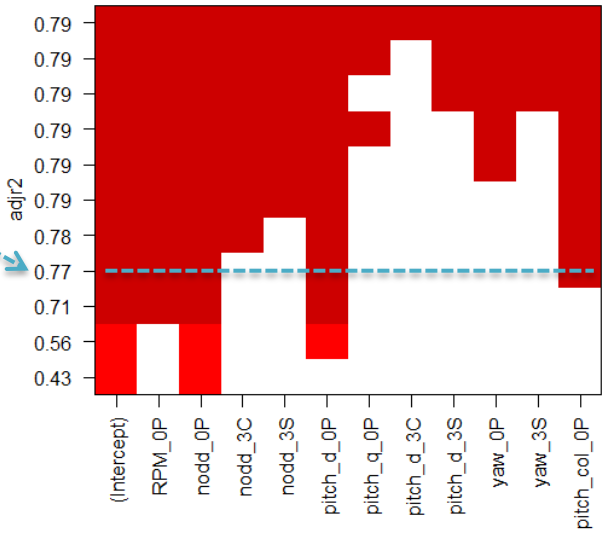


Figure 24 Best sub-model selection for linear regression. x-axis is predictor name. y-axis is the highest adjusted R^2 value for sub-model with same number of predictors. The color box indictes if the predictor is selected in the sub-model

Figure 23 showed that adjusted R^2 improves significantly within the first 4 predictors. Figure 24 showed predictor names used in each best sub-model. For example, for sub-model with 4 predictors, the algorithm compared all possible 4-predictor combinations in 11 total

predictors and down-selected the sub-model with highest adjusted R^2 of 0.77 (highlighted by the arrow in Figure 23). The corresponding 4 predictors are "RPM_0P", "nodd_0P", "pitch_d_0P" and "pitch_col_0P" (highlighted by horizontal line in Figure 24).

4.2.2 Hyperparameters

Following the same process as classification models in Part A, regression models with CART, Neural Network, SVM, Random Forest and XGB were developed. Details of hyperparameters optimization using a grid search were shown in Figure 38 (Appendix D) for RMSE metric. Optimized model hyperparameters from 5-repeat of 10-fold CV process were summarized in Table 11.

Table 11 Optimized hyperparameters for regression models (Part B)

Model Name	Hyper Parameters
Generalized Linear Model (GLM)	alpha = 0.5, lambda = 3.793e-4
Classification and Regression Tree (CART)	cp = 1.987e-4
Neural Network (NN)	size = 11, decay = 0.01
Support Vector Machine (SVM)	sigma = 0.148 and C = 4
Random Forest (RF)	mtry = 6
eXtreme Gradient Boosting (XGB)	nrounds = 1000, max_depth = 10, eta = 0.1, gamma = 0,

4.2.3 Model Performance

Two classification models can be visualized. Figure 25 showed the CART regression tree structure with the optimized hyperparameter. The root predictor is "Pitch_d_0P". "nodd_0P" is the second level splitter. These are different from classification CART tree in Figure 17.

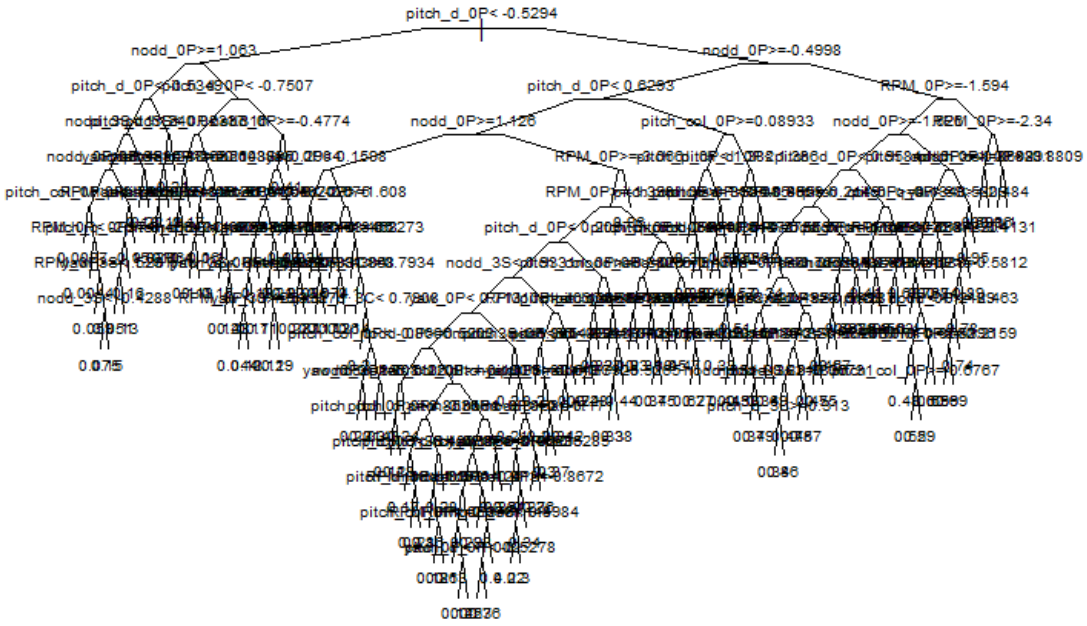


Figure 25 CART regression model (Part B).

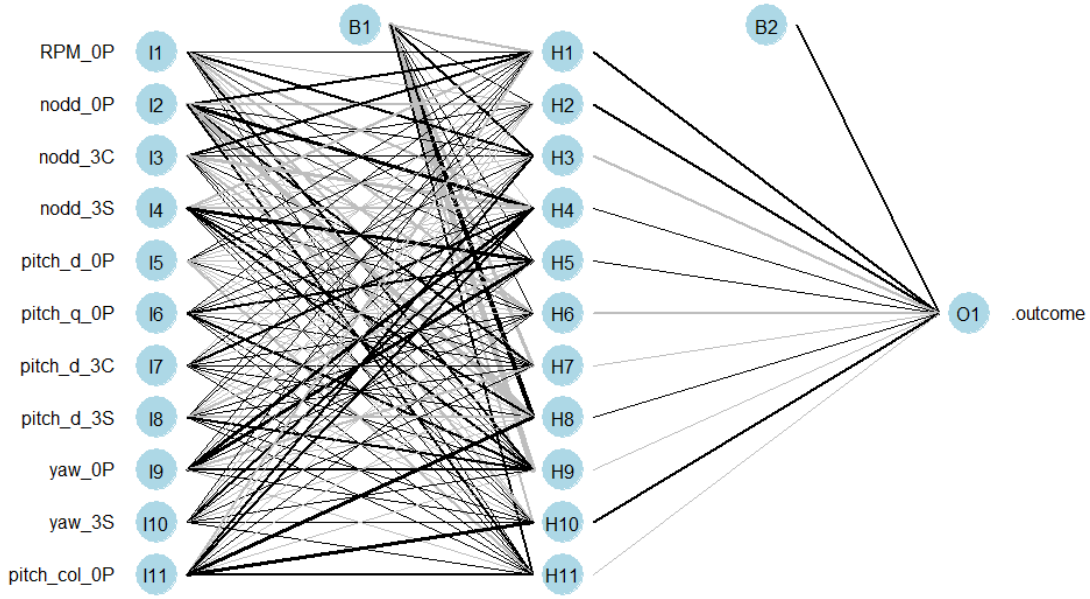


Figure 26 Neural Network regression model in Part B.

Figure 26 showed optimized Neural Network model with 1 hidden layer (11 units) and 1 regression output. Figure 27 showed predictor importance for Random Forest model. Here, the y-axis is the predictor's impact on the model's MSE. Top 3 predictors are "pitch_d_0P", "nodd_0P" and "RPM_0P", which are different from top 3 ones from classification Random Forest model in Figure 19. However, "nodd_P" is important for both models.

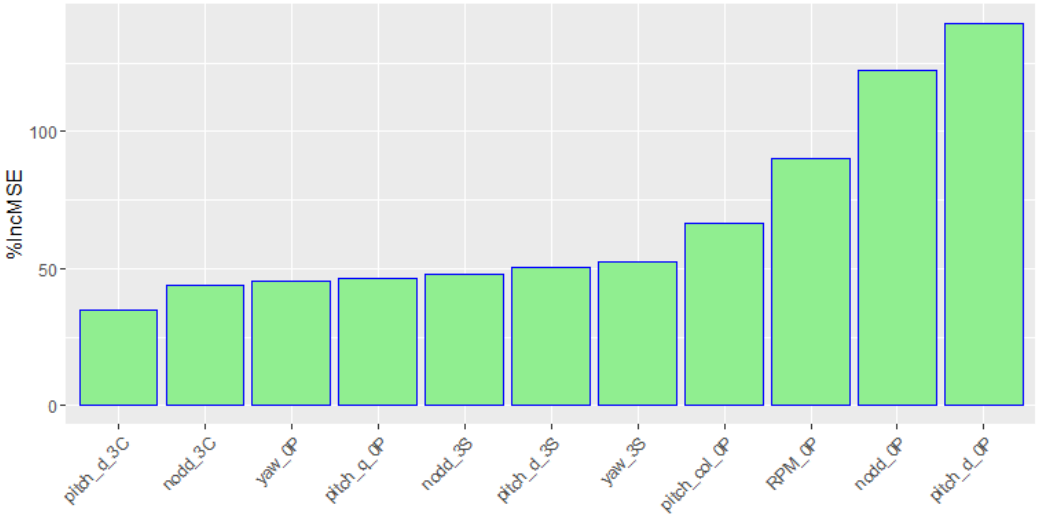


Figure 27 Variable importance from Random Forest regression model

To check if there are any residual patterns in predictions from regression models using test dataset, Figure 28 compared results from GLM (blue) and XGB (red). x-axis is true alpha calculated from Eq.(1) and y-axes are predicted alpha (left) and residuals (right) respectively. The advantage of XGB model is clearly shown. Residual plots for other regression models were summarized in Figure 39 (Appendix E). Only CART model showed some patterns.

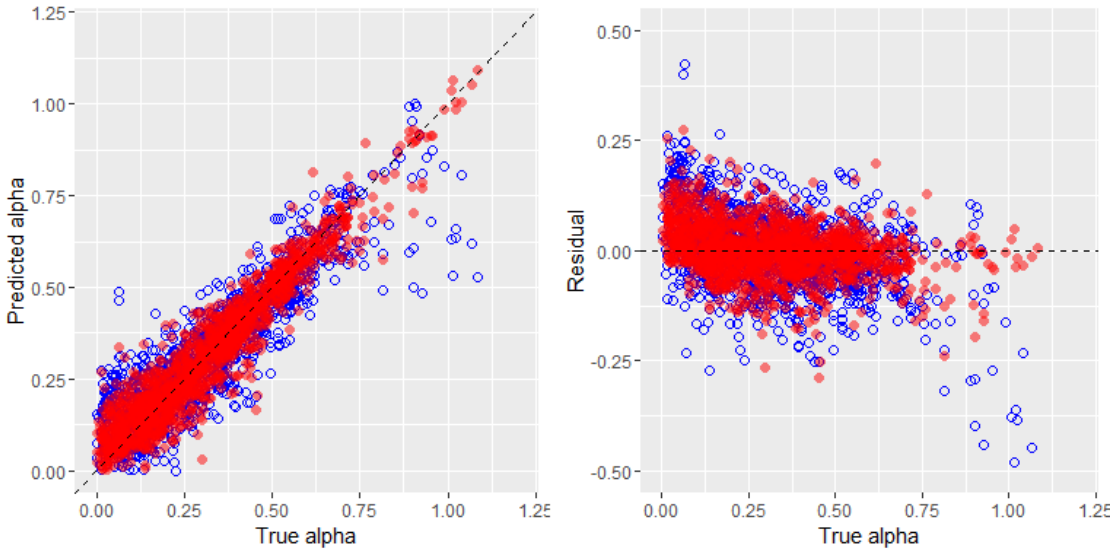


Figure 28 Predicted alpha and corresponding residuals for regression models of GLM (blue) and XGB (red) using test data. X-axis is true alpha from Eq.(1).

To evaluate model prediction performance, one approach is to check RMSE and R^2 metric variations from the 5-repaet 10-fold CV process as in Figure 29. It clearly showed metric variations from the training process of 50 runs. The Random Forest and XGB models have better metrics than the rest. GLM model is still the worst.

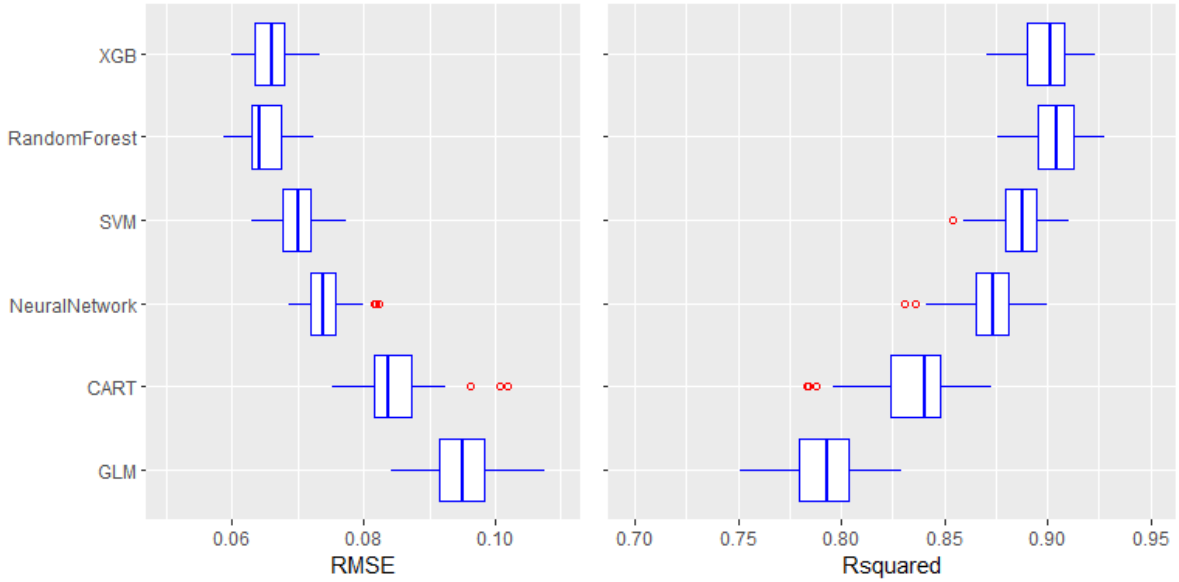


Figure 29 Regression model comparison with RMSE and R^2 metrics using training data (Part B). Model with lower RMSE and higher R^2 is better.

Regression model performance were summarized in Table 12 with metrics of RMSE and R^2. These performance metrics were evaluated based on the same test dataset for consistency. It should be noted that these model performances are little different from Figure 29. Figure 29 were created with training dataset from the 5-repeat of 10-fold CV process and Table 12 is from test data.

Table 12 Summary of regression model performances with test data (Part B).

Parameter	GLM	CART	Neural Network	SVM	Random Forest	XGB
RMSE	0.0929	0.0802	0.0733	0.0657	0.0623	0.0603
R^2	0.800	0.851	0.875	0.900	0.911	0.916

The final model ranking based on Table 12 with RMSE metric is:

```
XGB > Random Forest > SVM > Neural Network > CART > GLM
```

Regression model ranking is similar to classification ranking in Part A. XGB and Random Forest are still the top two models and GLM is of the least performance in both parts.

4.2.4 Impact of Filtered Predictors

Following the same process in Section 4.1.5, Table 13 compared RMSE and R^2 metrics of GLM-all and XGB-all models (14 predictors) to GLM and XGB (11 predictors). The hyperparameters of these two models were re-optimized based on the RMSE metric. Table 13 indicated no impact on regression model performance due to filtered out predictors.

Table 13 Impacts of filtered-out predictors on regression models (Part B). GLM-all and XGB-all models used 14 predictors, and GLM and XGB used 11 filtered predictors.

Parameter	GLM	GLM-all	XGB	XGB-all
RMSE	0.0929	0.0923	0.0603	0.0602
R^2	0.800	0.802	0.916	0.916

4.2.5 Wind Speed Profiles

The wind shear is calculated from wind speed profile by definition. It is always prudent to verify fitted wind speed profiles using model parameter alpha predicted by XGB model.

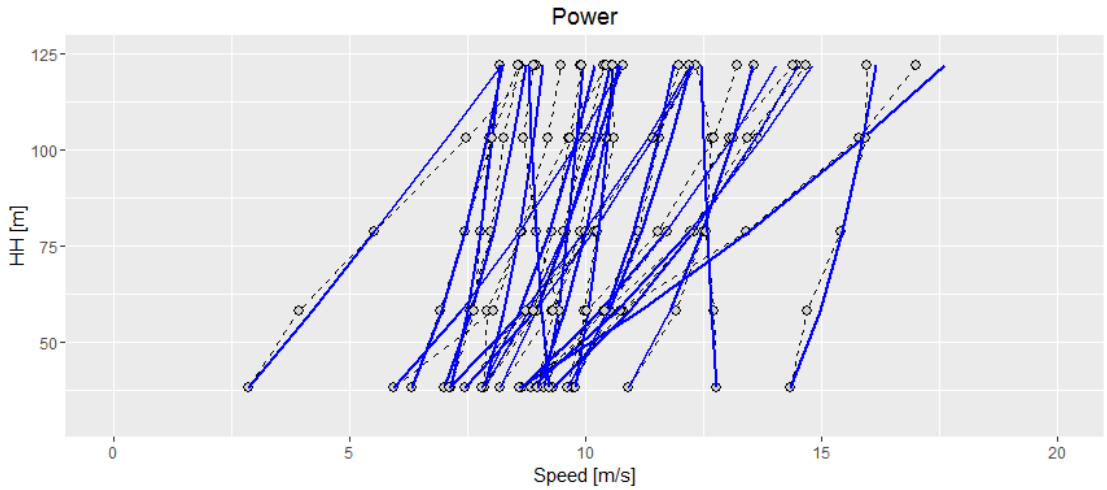


Figure 30 Comparing true and fitted wind speed profiles. 24 profiles were randomly selected as in Figure 9. Fitted profiles were created using alpha predicted by XGB model.

Figure 30 compared true wind speed profiles from provided dataset (black dot lines with symbol) to fitted profiles (blue lines). 24 profiles were randomly selected as in Figure 9. Fitted profiles matched most of the true ones. There are some discrepancies, especially at high altitude of 122m. Field engine need to evaluate whether these discrepancies are important or not for wind shear estimation.

4.3 Part C - Estimation of Wind Speed Profile for Other Classes

In Part C, regression models for wind shear classes of LLJ and Flat were developed. The Other profile was ignored due to no clear profile pattern and only 129 samples. Also, only GLM and XGB models were considered as the worst and best model based on results in Part B.

To estimate the wind speed profile, following process were developed

- Develop different wind speed profile model for each wind shear class.
- Estimate wind speed profile model parameters for each wind shear class using wind speed sensor data.
- Create regression models using load senor data (14 predictors) to estimate model parameters for each class.

4.3.2 LLJ Wind Speed Profile

For the LLJ wind speed profile (Figure 9), the wind speed profile has a large curvature and smaller wind speed at the top. This profile can be modeled as a polynomial curve:

$$V(h) = c0 + c1 * HH + c2 * HH^2 \quad (3)$$

coefficients of c0, c1 and c2 can be calculated by fitting at least 3 wind speed sensor data, which may not be practical for field wind turbines. Models with 2 sensors were also considered, but it is hard to capture high profile curvature, especially at high altitude.

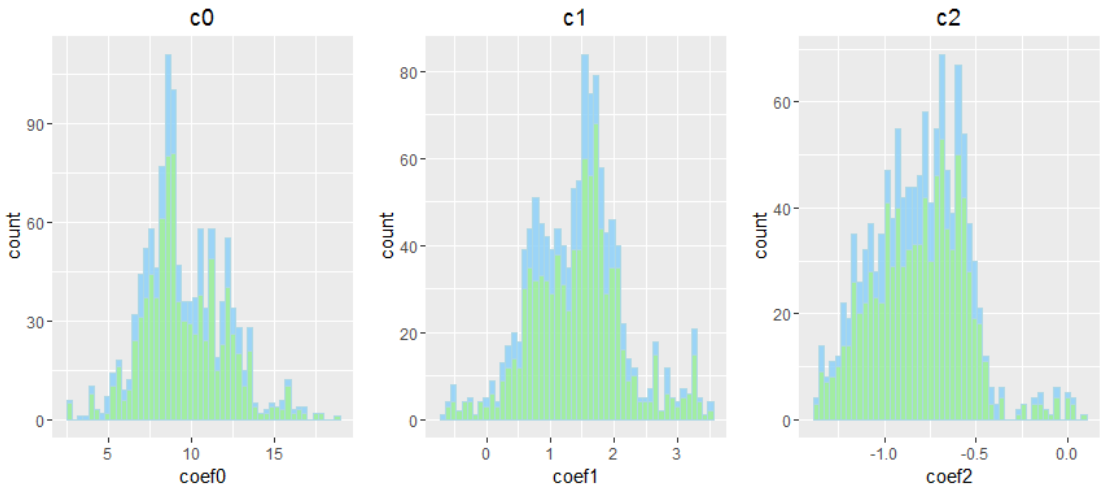


Figure 31 Split of training and test data for model parameter c0, c1 and c2 with LLJ class (Part C). Green is for training data and blue is for test data.

Figure 31 showed distributions of calculated c0, c1 and c2 coefficients for LLJ class. The split of training and test data was also illustrated. To estimate coefficients of c0, c1 and c2, altitude (HH) was centered and scaled. Three regression models were developed to fit each coefficient. Both GLM and XGB models were optimized and detail training process was shown in Figure 40 (Appendix F) using RMSE metric.

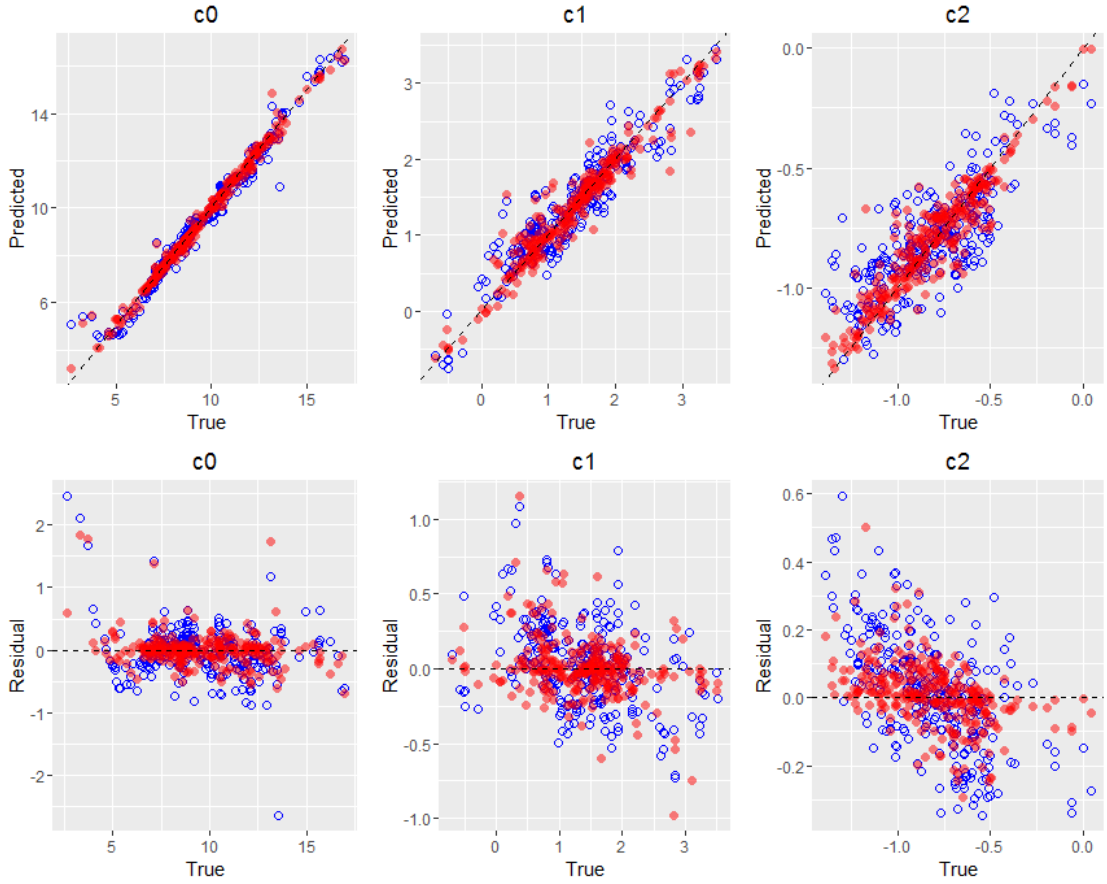


Figure 32 Predicted (top) model parameters c0, c1 and c2, and corresponding residuals (bottom) for regression models of GLM (blue) and XGB (red) using test data for LLJ class (Part C). X-axis is true model parameters c0, c1 and c2 in Eq.(3).

Predicted coefficients and residuals from both GLM and XGB models were shown in Figure 32. There is no clear pattern in residual plots. Model performance metrics of RMSE and R^2 were summarized in Table 14. It can be seen clearly that both GLM and XGM model predicted c0 very well with R^2 close to 1. However, model performance degraded for coefficient c2. For all three coefficient predictions, XGB models are consistently better than corresponding GLM models. These agreed with results in Part B (section 4.1)

Table 14 Summary of regression model performances for LLJ profile (Part C).

Model	GLM			XGB		
Parameter	c0	c1	c2	c0	c1	c2
RMSE	0.404	0.270	0.165	0.273	0.201	0.096
R^2	0.975	0.874	0.597	0.989	0.930	0.864

Figure 33 compared true wind speed profiles from provided dataset (black dot lines with symbol) to fitted profiles (blue lines). 24 profiles were randomly selected as in Figure 9. Fitted profiles were created using the predicted coefficients c0, c1 and c2 from XGB models and Eq.(3). Fitted profiles matched well to true profiles.

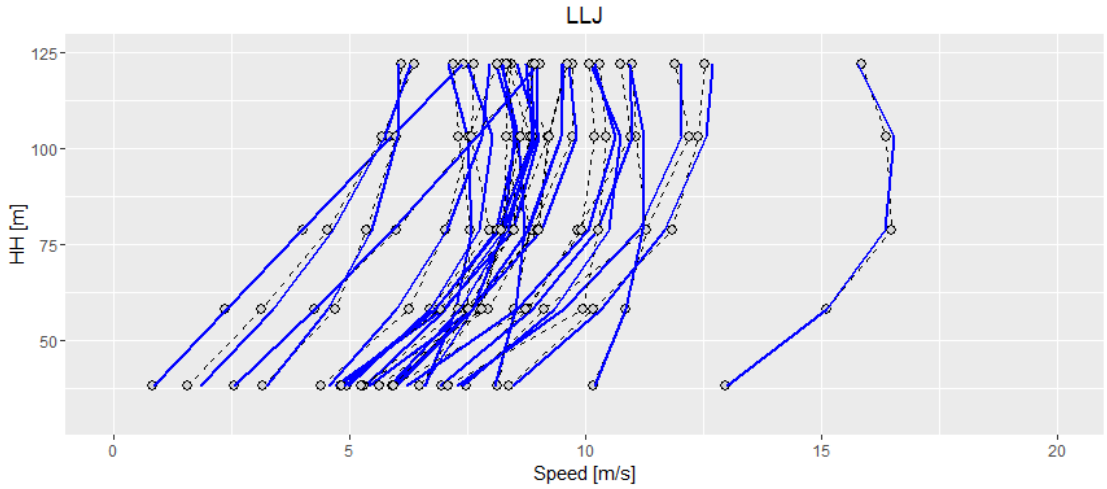


Figure 33 Comparing true and fitted wind speed profiles. 24 profiles are randomly selected as in Figure 9. Fitted profiles were created using c0, c1 and c2 predicted by XGB model.

4.3.2 Flat Wind Speed Profile

For the Flat wind speed profile (Figure 9), the speed profile can be modeled as:

$$V(h) = slope * (\mathrm{HH} - 38) + V_{38m} \tag{4}$$

the slope is calculated by using wind speeds at altitude of 78.7m as,

$$slope = (V_{78.7m} - V_{38m})/(78.7 - 38) \tag{5}$$

Figure 34 showed slope distribution. The split of training and test data is also illustrated.

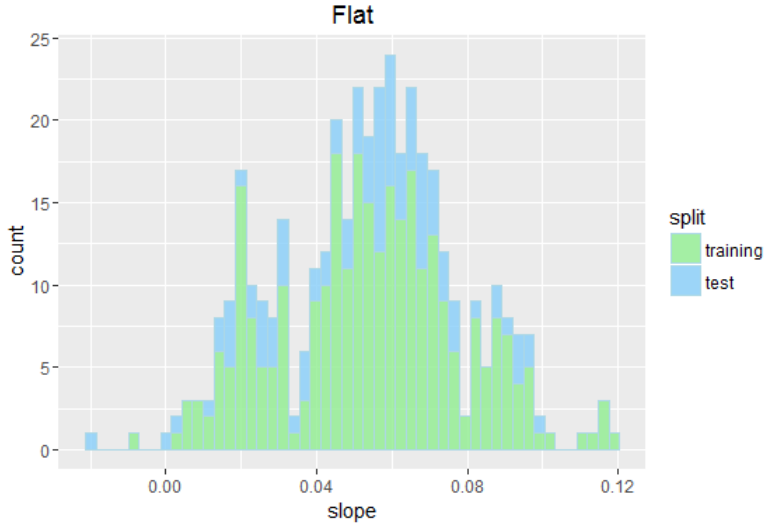


Figure 34 Split of training and test data for model parameter slope with Flat class (Part C).

Both GLM and XGB models were optimized and the training process was shown in Figure 41 (Appendix G) using the RMSE metric. Predicted slopes and residuals from both GLM and XGB models were shown in Figure 35. There are no clear patterns in residual plot.

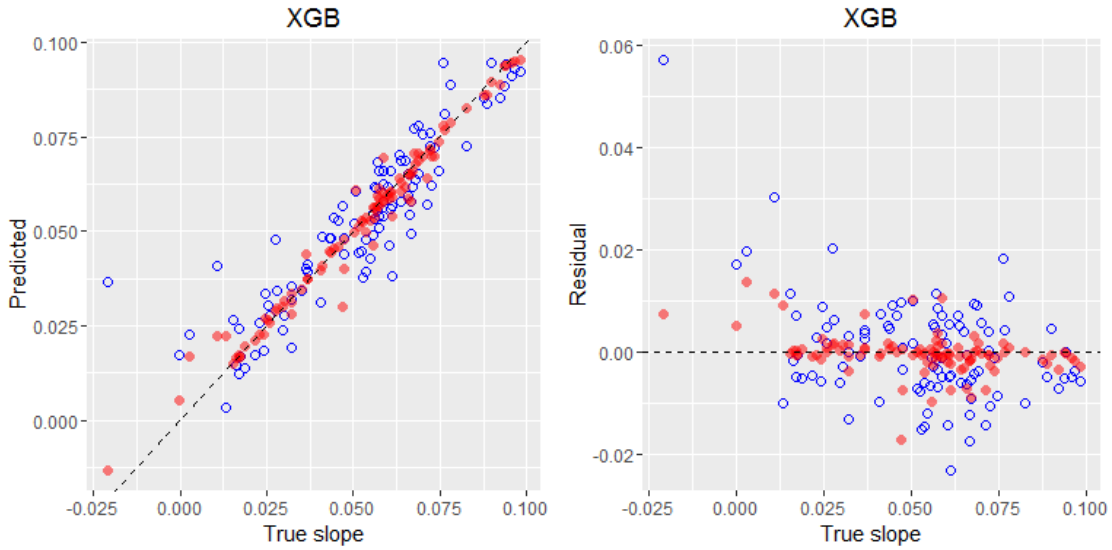


Figure 35 Predicted slopes and corresponding residuals for regression models of GLM (blue) and XGB (red) using test data for Flat class (Part C). X-axis is true slope calculated from Eq.(4).

Table 15 compared RMSE and R^2 metrics using test dataset. The advantage of XGB model is clearly shown by Figure 35 and Table 15.

Table 15 Summary of regression model performances for Flat class (Part C).

Parameter	GLM	XGB
RMSE	0.0104	0.00407
R^2	0.805	0.972

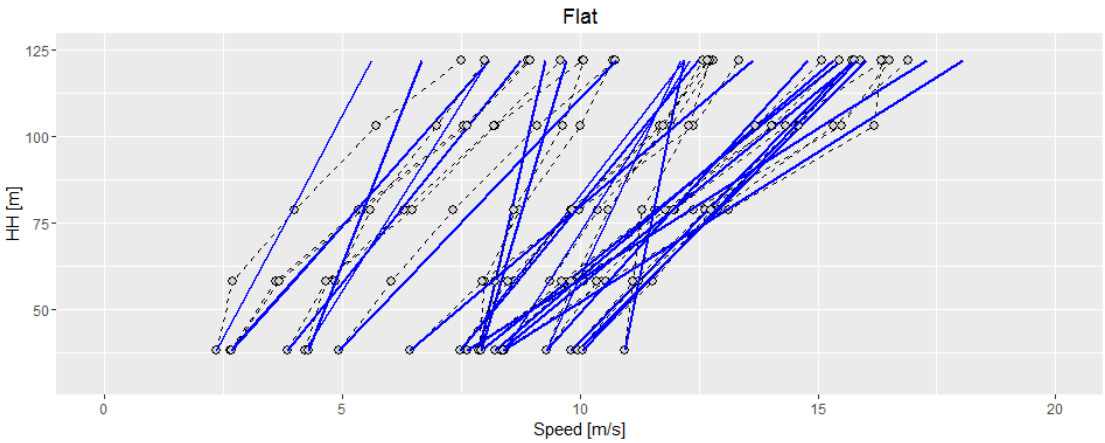


Figure 36 Comparing true and fitted wind speed profiles. 24 profiles are randomly selected as in Figure 9. Fitted profiles were created using model parameter slope predicted by XGB model.

It should be noted that total samples in Flat class is only 394. For this limited sample size, prediction model performance evaluation may not be accurate. For example, the R^2 value of XGB is 0.97 which may seem too high based on previous results in Part B. Collecting more data is the Flat class will help evaluating model performances.

Figure 36compared true profiles from provided dataset (black dot lines with symbol) to fitted profiles (blue lines). 24 profiles were randomly selected as in Figure 9. The fitted profiles were created using predicted slopes from XGB model and Eq.(5). The fitted profiles matched most of true profiles at lower altitude. Bigger discrepancies were observed at 122m altitude. A better wind speed profile model, like polynomial model in Eq.(3), will help.

5 Conclusions

In this case study, supervised machine learning models of General Linear Models (GLM), Classification and Regression Tree (CART), Neural Network (NN), Support Vector Machine (SVM), Random Forrest (RF) and eXtreme Gradient boosting (XGB) were explored. These models can be used for both classification and regression problems.

In part A, models have been built to classify samples into four wind shear classes. Due to imbalanced data distribution, the Baseline Accuracy (0.79) is very high. Model with best prediction performance is XGB with Kappa=0.74 (Accuracy=0.92). The least accurate model is GLM with Kappa=0.49 (Accuracy=0.85). Classification model ranking based on Kappa metric from test data is

```
XGB > Random Forest > SVM > CART > Neural Network > GLM
```

In part B, models have been developed to estimate wind shear (alpha) for wind speed profile of Power Law only. Standard linear model, linear model with best sub-model selection and GLM all have similar model performance with RMSE=0.093 (R^2=0.80). Linear models are the least accurate in part B. Model with best prediction performance is XGB with RMSE=0.60 (R^2=0.92). Regression model ranking based on RMSE from test data is

```
XGB > Random Forest > SVM > Neural Network > CART > GLM
```

To estimate wind shear in Part C, polynomial and linear wind speed profile models were developed for LLJ and Flat class, respectively. GLM and XGB models were then developed to predict wind speed profile parameters. XGB model consistently outperformed GLM model.

The current case study presented **satisfactory results** for both classification (Part A) and regression (Part B & C) problems with down-selected XGB model. However, it is important to note assumptions made in above analysis. Following issues should be considered.

- **Confounding predictors (variables).** For wind turbine operation, environmental factors like wind speed, temperature, altitude, weather conditions, wind turbine locations etc. can affect wind speed profile and shear.

  Furthermore, the provided data may be only for one specific wind turbine model at one particular location. Data from other wind turbine models and locations should be

included to further evaluate the down-selected model performance. It is highly possible that each wind turbine model at each site will have its own prediction model.

- **Data imbalance.** As can be seen in Figure 6 to Figure 8, the provided data set only includes continuous sampling data during the 12-day measurements at 2 min interval. Especially, Power Law class is dominant in the first 9 days and major class is LLJ in the last 3 days. Data samples for wind speed classes of Flat and Other are really small. This leads to very high baseline accuracy for wind shear classification and not enough data for regression in Part C for minority classes. The physical reasoning for data distribution should be considered. More data in other wind speed classes are needed for better prediction accuracy, especially for minority classes.

- **Data quality.** One critical missing part in the provided dataset is measurement uncertainty for load sensors. Measurement uncertainty can potentially affect model training and down-selected learning models.

  Potential outlier data were shown in Figure 4. Whether these outlier data can be filtered or not, have to be judged by the field engineer. Removing outlier data has the potential to further improve model prediction performance.

- **Wind shear estimation.** In the current case study, wind shear (alpha) calculation was provided in Eq.(1) for Power Law by Xu and Evans (2016). Models for ShearTypeClass of LLJ, Flat and Other were not provided. There may be better prediction models if wind profiles were used as the target. Also, overall model performance should include combined effects of classification errors in Part A and regression errors in Part B and C. These will not change the model ranking, but model performances will be different.

- **Time series data**. The provided dataset is time-series in nature. The current case study was limited to load sensor data (14 predictors) only. However, temporal coherence can be leveraged to enhance model prediction accuracy. For example, if data can be split into two time windows with the dividing line at 2015-06-26, the baseline accuracy will be 0.96 for data before 2015-06-26.

- **Risk and Financial Impacts.** Besides technical aspects of data analysis, it is always important to include domain experts to understand business values and financial risks. The down-selected model may have to balance technical metric (Kappa or RMSE) and financial risks. For example, if Flat wind speed profile is a worst case and can potentially destroy the wind turbine, a model should be selected to have high classification performance for this class. Here, training metrics with weighting factors can help improving minority class prediction performance.

Even though the current case study showing satisfactory results, I would **NOT recommend** GE Wind to launch "Virtual Wind Shear Sensor" program at this time. Extended sample size, balanced dataset, added confounding factors, and risks/financial analysis must be considered. A **pilot verification** program is recommended for different wind turbine models at different sites.

References

1. Xu Yunwen and Scott Evans, "GE Analytics Engineer Program Case Study: Virtual Wind Shear Sensor." GE internal, 2016. http://libraries.ge.com/foldersIndex.do?entity_id=36905398101&sid=101&SF=1.

2. Gareth James, Witten Daniela, Hastie Trevor, Tibshirani Robert, "An Introduction to Statistical Learning with Applications in R", Springer, 2015. http://www-bcf.usc.edu/~gareth/ISL/

3. Kuhn Max, "Never Tell Me the Odds! Machine Learning with Class Imbalances", 2016, http://schedule.user2016.org/event/7Bac/never-tell-me-the-odds-machine-learning-with-class-imbalances-part-1

4. Kuhn Max and Johnson Kjell, "Applied Predictive Modeling", 2013, Springer http://appliedpredictivemodeling.com/

5. Kuhn Max, "Building Predictive Models in R Using the caret Package", Journal of Statistical Software, Vol.28, Issue 5, November 2008 https://www.jstatsoft.org/article/view/v028i05

6. Fernandez-Delgado Manuel, Cernadas Eva, Barro Senen, Amorim Dinani, "Do we Need Hundreds of Classiers to Solve Real World Classification Problems?", Journal of Machine Learning Research, 2014, Vol.15, pp.3133-3181 http://www.jmlr.org/papers/v15/delgado14a.html

7. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., "SMOTE: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, 16:321-357, 2002 http://jair.org/media/953/live-953-2037-jair.pdf

Appendix A - Data Description

The data was provided in file "WindFarm_2min_AnalyticsEngineer_wWS.csv". Wind speed from a 5 sensor met mast were described in Table 16. It should be noticed that m58, m102 and m122 are not available for most installed wind turbines. Load sensor data from turbine sensors (14 parameters) were shown in Table 17. Wind speed profiles were labeled by field engineers as Power Law (0), LLJ (1), Flat (2), and Other (3), as shown in Table 18.

Table 16 Description of wind speed sensor parameters

Name	Description
m38	wind speed at height of 38 m, [m/sec]
m58	wind speed at height of 58 m, [m/sec]
m78	wind speed at height of 78 m, [m/sec]
m103	wind speed at height of 103 m, [m/sec]
m122	wind speed at height of 122 m, [m/sec]

Table 17 Description of wind turbine load sensor parameters (X variables)

Name	Description
RPM_0P	0P component of RPM, [1/min]
nodd_0P	0P component of nodding moment, [Nm]
nodd_3C	3P (cosine) component of nodding moment, [Nm]
nodd_3S	3P (sine) component of nodding moment, [Nm]
pitch_d_0P	0P component of d-component of pitch angle, [deg]
pitch_q_0P	0P component of q-component of pitch angle, [deg]
pitch_d_3C	3P (cosine) component of d-component of pitch angle, [deg]
pitch_d_3S	3P (sine) component of d-component of pitch angle, [deg]
yaw_0P	0P component of yawing moment, [Nm]
yaw_3C	3P (cosine) component of yawing moment, [Nm]
yaw_3S	3P (sine) component of yawing moment, [Nm]
P_el	0P component of electrical power, [kW]
V_estim	0P component of MBC estimated wind speed, [m/s]
pitch_col_0P	0P component of collective pitch angle, [deg]

Table 18 ShearTypeClass

ShearTypeClass	Wind Speed Profile
0	Power Law
1	LLJ (Low-Level-Jet)
2	Flat
3	Other

Appendix B - Classification Model Training (Part A)

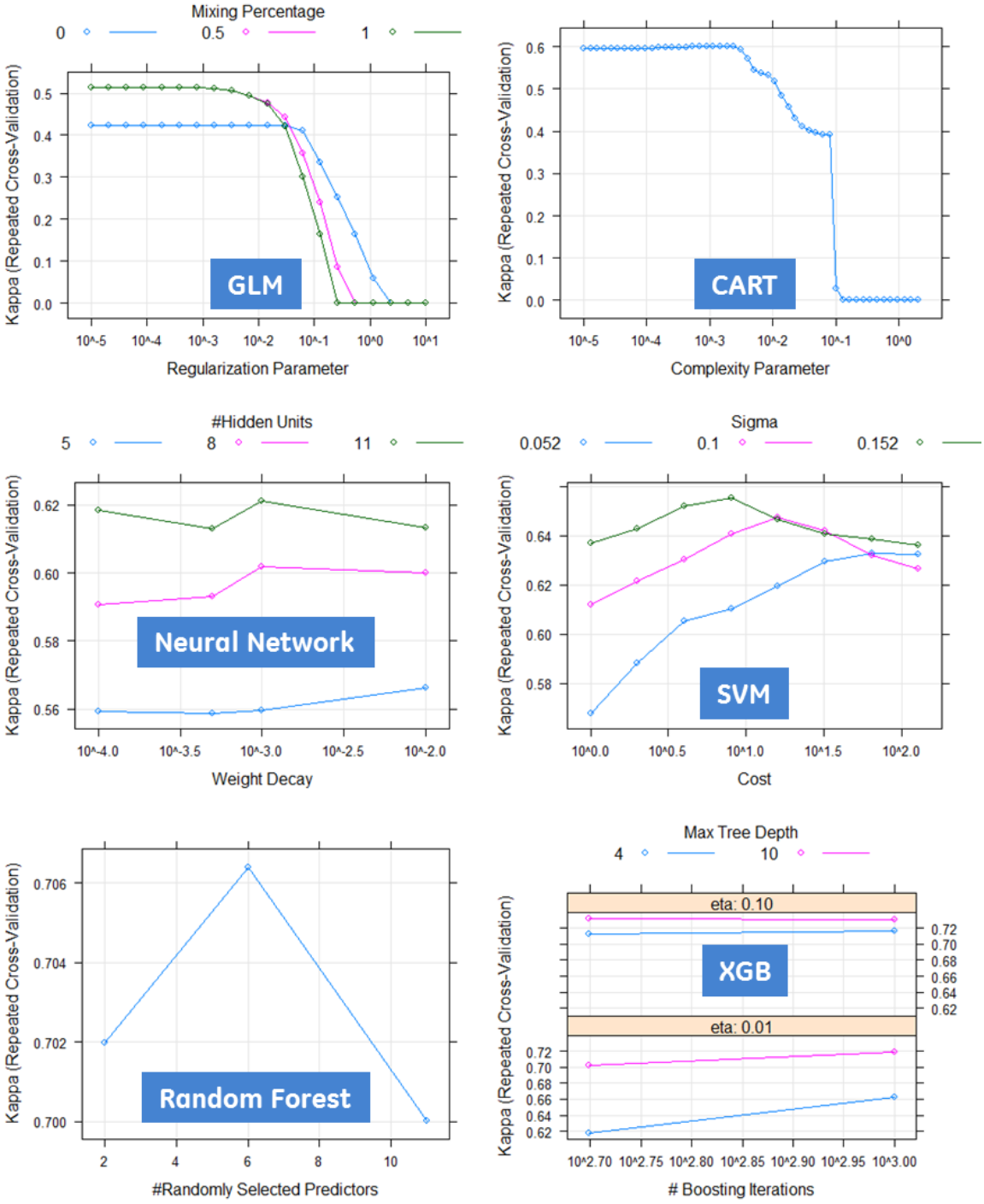


*Figure 37 Training process for classification models using **Kappa** metric (Part A). Model with higher Kappa is better.*

Appendix C - Linear Regression (Part B)

Outputs from standard linear regression was shown in Table 19. Fitting coefficient and p-value for each predictor were included. Fitting coefficients are for predictors in raw units.

Table 19 Linear regression model with 11 filtered predictors (Part B).

```
## Call:
## lm(formula = alpha ~ ., data = wind.sub0[tr0, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44131 -0.05189  0.00052  0.04927  0.60816
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.866e-01  1.601e-02  36.647  < 2e-16 ***
## RPM_0P       -4.520e-02  1.292e-03 -34.981  < 2e-16 ***
## nodd_0P      -4.853e-04  6.330e-06 -76.670  < 2e-16 ***
## nodd_3C       2.693e-04  3.001e-05   8.975  < 2e-16 ***
## nodd_3S       1.903e-04  2.229e-05   8.540  < 2e-16 ***
## pitch_d_0P    1.404e-01  2.115e-03  66.367  < 2e-16 ***
## pitch_q_0P   -3.423e-02  5.990e-03  -5.714 1.17e-08 ***
## pitch_d_3C    3.544e-02  2.034e-02   1.742   0.0816 .
## pitch_d_3S   -2.023e-01  3.102e-02  -6.522 7.67e-11 ***
## yaw_0P        1.099e-04  1.912e-05   5.745 9.79e-09 ***
## yaw_3S       -2.514e-04  4.212e-05  -5.969 2.57e-09 ***
## pitch_col_0P -1.586e-02  6.198e-04 -25.582  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09486 on 4747 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7917
## F-statistic:  1645 on 11 and 4747 DF,  p-value: < 2.2e-16
```

Appendix D – Regression Model Training for Power Law (Part B)

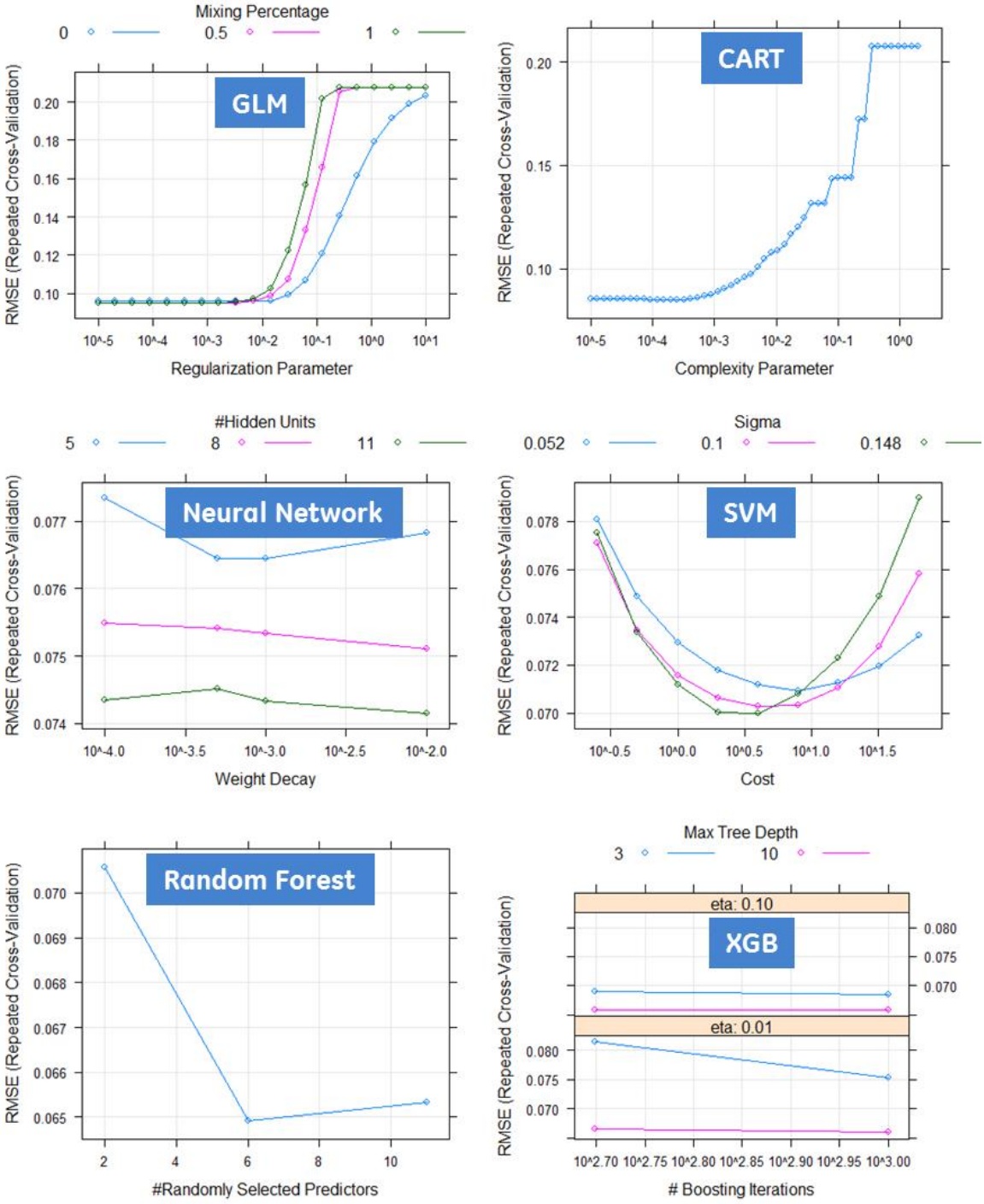


*Figure 38 Training process for regression models using **RMSE** metric (Part B). Model with lower RMSE is better.*

Appendix E – Residual Plots (Part B)

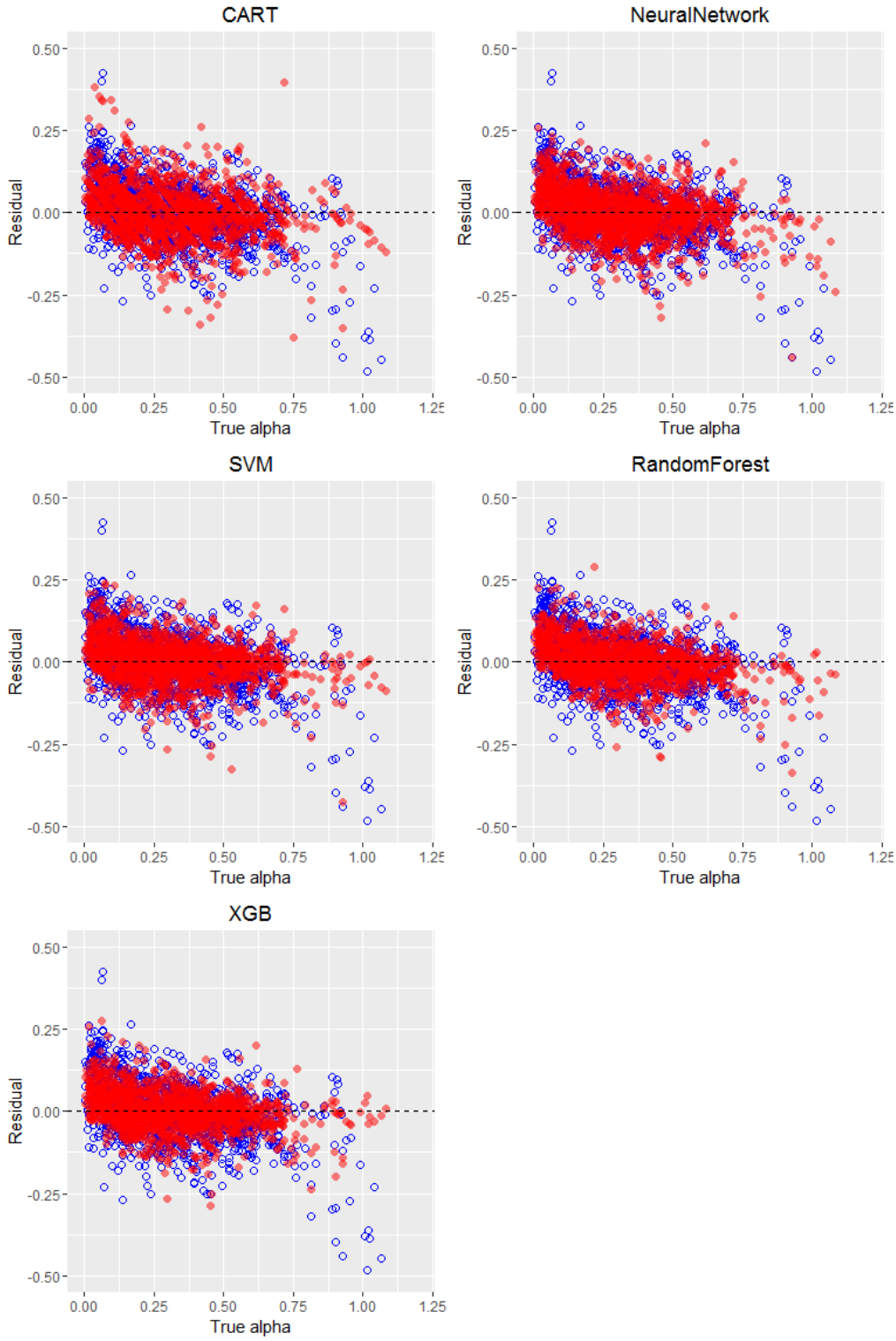


Figure 39 Residuals plots for regression models using test data (Part B). X-axis is true alpha calculated from Eq.(1). Y-axis is the difference between predicted and true alpha.

Appendix F – Regression Model Training for LLJ Profile (Part C)

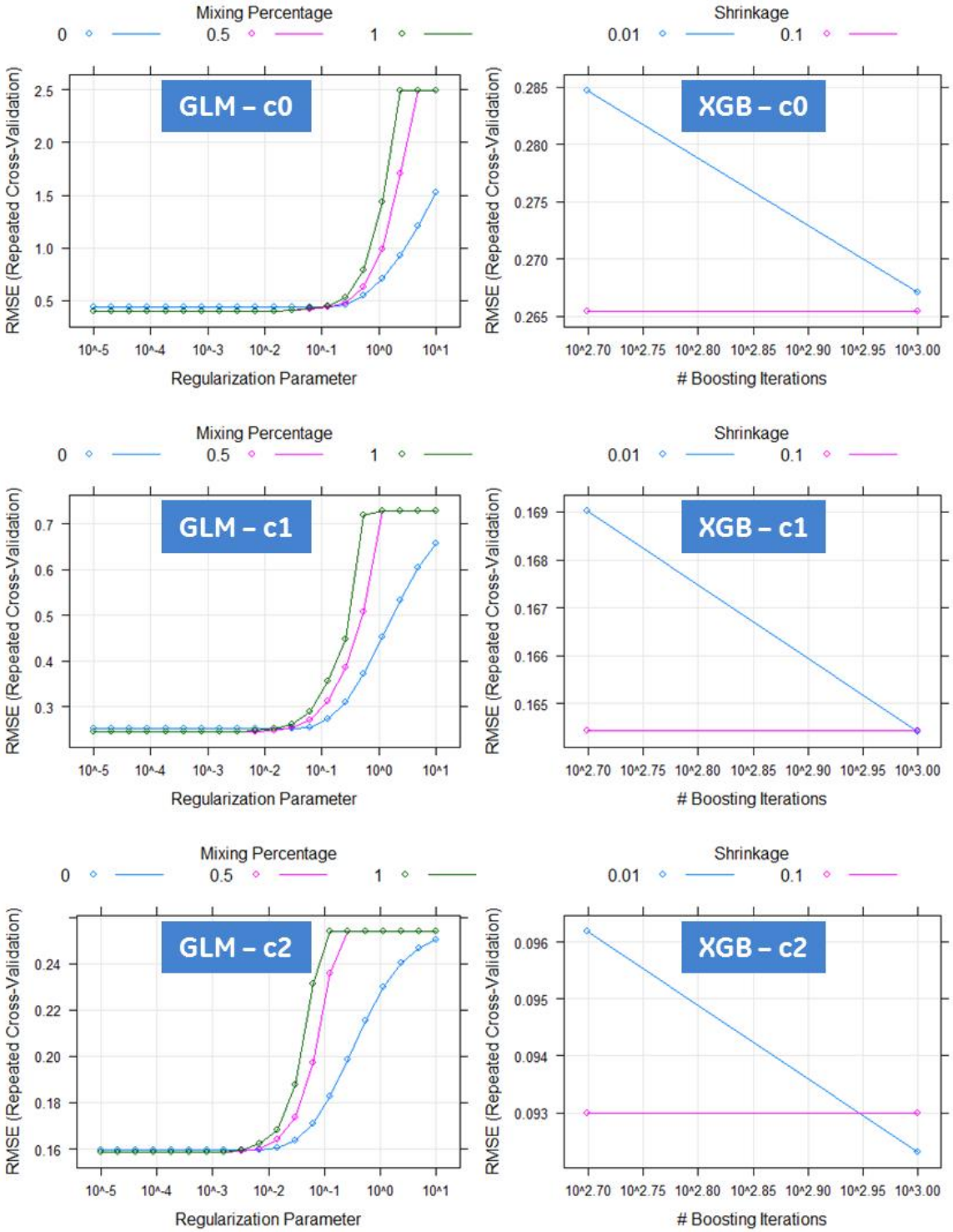


*Figure 40 Training process for GLM and XGB regression models using **RMSE** metric (Part C, ShearTypeClass = LLJ). Each coefficient c0, c1 and c2 has its own model. Model with lower RMSE is better.*

Appendix G – Regression Model Training for Flat Profile (Part C)

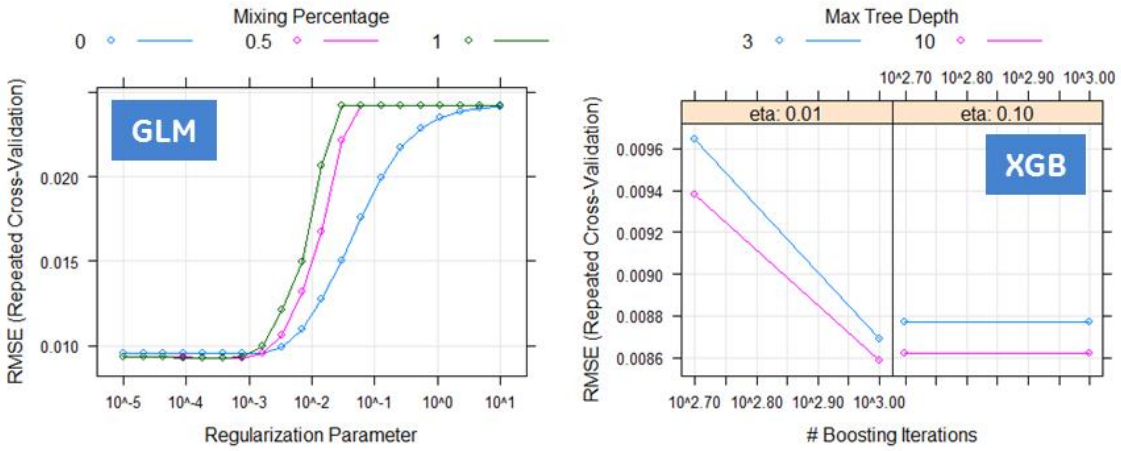


*Figure 41 Training process for GLM and XGB regression model using **RMSE** metric (Part C, ShearTypeClass = Flat). Model with lower RMSE is better.*