STAT 440 MODULE 1 REPORT

PREDICTING ARCTIC UNDERGROUND TEMPERATURES

October 3, 2018

Xuankang Zhu 301317536

Anson Hsu 301208466

JiaJie Lim 301321990

02 October 2018

Contents

Introduction

Our goal in this module is to predict the temperature of soil at a depth of 100cm in approximately 80 distinct Arctic locations between the year 1998 and 2016. We were given a training set with 2,748 samples and a test set with 660 samples. Our first approach was to perform data augmentation, by making minor alterations to our existing dataset to increase the amount of relevant data. In principle, it is said that the more the data, the better our models will be. Then, we applied linear regression and KNN imputation for missing values to draw a better inference accuracy of the dataset. Subsequently, we presented four different models to allow the production of better predictive performance compared to a single model. Cross-validation was applied to access the predictive performance of our models and to verify if the model is overfitting or underfitting. Then, we attempted ensemble learning by combining our final predictions from four different models and computed the weighted average mean. We then successfully obtained a low RMSE which is approximately 1.1 in our test set.

Data description

This data is about 3,408 observations of permafrost underground temperature readings which were sampled at a depth of 100 cm in around 80 distinct Arctic locations between the years 1998 and 2016. The explanatory variables were introduced in module 1 hand papers.

Data examination

By examining the summary of the datasets, we found that some observations (train and test sets) have one or multiple missing variables. We plotted the density distribution graphs(figure 1) of the response variable(Temp100cm) respected to variable-missed entries and completed entries, and concluded that the variables are missing at random. Subsequently, we looked at the paired scatter plots(figure 2) of the data and found that our response variable is strongly correlated to TempAir, TempSurf and Month. Moreover, since the Temp100cm vs Month scatter plot is a bell-curved and Month variable has only 12 different entries, we convert the Month variable into a categorical variable. The facet-separated scatter plots(figure 3) respected to Month shows that the distribution of other variables within each month is similar.

In order to build a model to make predictions on the test data, we have to fill in all the missing entries in the training set and test set to make a better prediction. We first added additional informational variables to provide more details with the variable missed entries:

Categorical variables:(all with same levels: 0 and 1)

ct1: whether or not the SnowDepth variable entries contain missing data.
ct2: whether or not the TempSurf variable entries contain missing data.
ct2: whether or the TempAir variable entries contain missing data.
Numerical variables:
CMTemp100cm: mean average of the Temp100cm respected to each month.
CMTempAir: mean average of the TempAir respected to each month.
CMTempSurf: mean average of the TempSurf respected to each month.
(Values for the additional variables for the test data are assigned based on the train data)

MULTIPLE IMPUTATION FOR THE MISSING VALUE

We first merged the train and test set together and sorted all the non-missing entries out to build our linear regression model. The model used TempAir, TempSurf, Year, Month and their interactions as regressor variables to predict the SnowDepth variable, it has really small RMSE 0.08 and is not over fitting proved by cross-validation. Then we used this model to fill in the missing entries of SnowDepth, and the remain missing entries with k-nearest neighbours imputation from the R-package "DMwR". Finally, we return the train and test set by splitting them back to their original size of observations which were then ready for modeling.

MODELING

We choose the Ensemble methods to promote better result for our models. Ensemble method is a machine learning technique that combines multiple base models and produces one optimal predictive model. Linear Regression is the ensemble method that we have chosen to combine the models and predict using the optimal model. For this Linear Regression ensemble method, we applied 4 different methods (random forest regression, neural network regression, KNN imputation, and gradient boosting) on our training set to build up 4 different models. Next, we predict the 4 different y-value(Temp100cm) by using the data in the training set in 4 different model; and then perform the linear regression to produce our final optimal predictive model. Then, we predict our final result by using the testing data set on the final model.

Gradient Boosting

Gradient boosting is a boosting method that reduces the high variance of learners by averaging lots of models fitted on bootstrapped data samples generated with replacement from training data; therefore this avoids overfitting. For the parameters of the model, we chose n.tree = 10000 because the higher number of times of bootstrap had to be done for a better prediction and an interaction.depth of 24 because that gave us the best estimation. Our gradient boosting give us really good prediction with low RMSE. In table1, we found out that gradient Boosting affects the ensemble model the most because it gives the most accurate prediction.

Neural Network Regression

Neural network is a method that simulates the working of neuron in the human brain. For this method, it will increase the accuracy after certain iterations. It will start as an unstable model with bias and end up with low bias result by minimizing the error. For the parameters, we selected size=2 which is the hidden layer(iterations), and decay=0.0001 that gave us the best estimation of our result.

Random Forest Regression

Random forest model is like a decision tree and is able to decide where to split based on a random selection of features. Random forest regression is used when the data is complex and non-linear, which was suitable for our data. In table1, Random Forest Regression has the second greatest effect on the final ensemble method model which was able to produce a good estimation of our data.

KNN

KNN imputation is the method to fill in the missing value but also can use to predict the response variable like a regression model. In our model, we were predicting the response variable, Temp100cm by imputing the missing values using KNN.

Conclusion

Ensemble method is a powerful method to reduce the biases and increase the accuracy of our prediction. As the proverb runs, all the models are wrong, but the average of them is less wrong. During our testing process, the RMSE produced by the ensemble method is significantly higher than the individual model. Furthermore, we can still improve our prediction accuracy by fine-tuning our individual model.

APPENDIX

0.0.1 Model formula

Temp100cm Year+Month+SnowDepth+TempAir+TempSurf+ ct1+ct2+ct3+CMTemp100cm +CMTempAir+CMSnowDepth+ Month:(SnowDepth+TempAir+TempSurf+CMTemp100cm +CMTempAir+CMSnowDeth)+ Year:(SnowDepth+TempAir+TempSurf+CMTemp100cm +CMTempAir+CMSnowDepth)+ ct1:(SnowDepth+TempAir+TempSurf+CMTemp100cm +CMTempAir+CMSnowDepth)+ct2:(SnowDepth+TempAir+TempSurf+CMTemp100cm +CMTempAir+CMSnowDepth)+ct3:(SnowDepth+TempAir +TempSurf+CMTemp100cm +CMTempAir+CMSnowDepth)

0.0.2 Table 1

Table 1: Linear Regression Coefficients for Ensemble Method

	Intercept	KNN	Random Forest	Gradient Boosting	Neural Network
Coefficients	-0.01354	0.09251	-0.21418	0.84743	0.25242

0.0.3 Figure 1

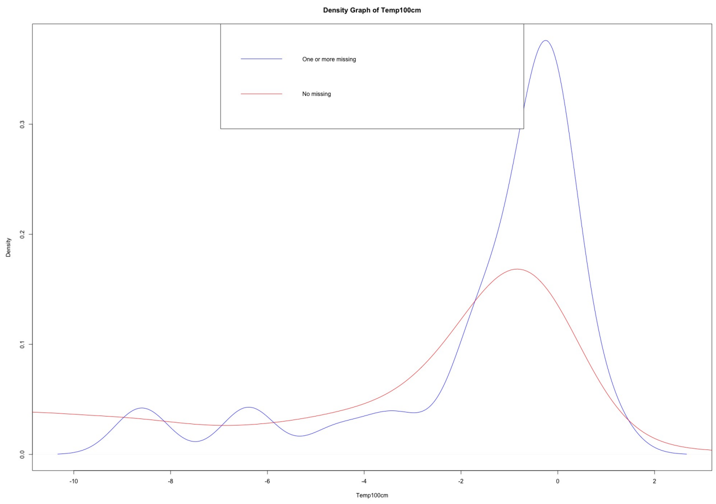

Figure 1: Density Distribution Plot

0.0.4 Figure 2

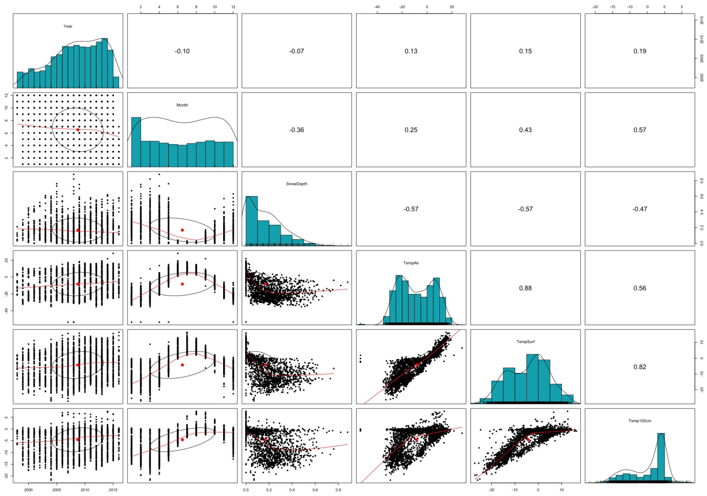

Figure 2: Paired Scatter plots

0.0.5 Figure 3

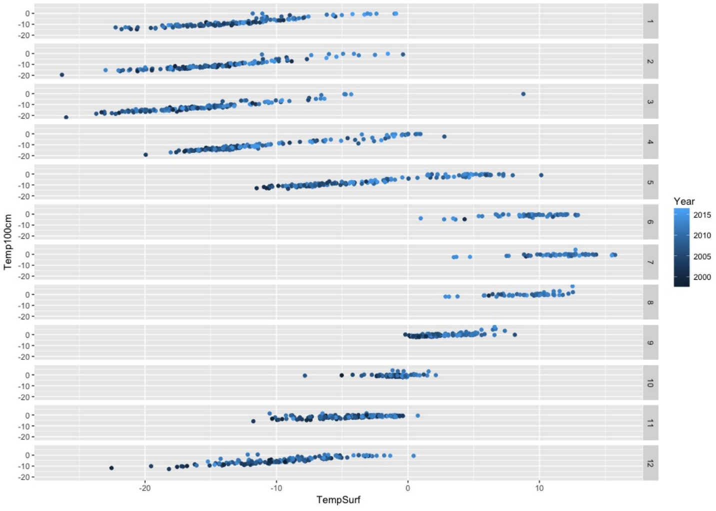

Figure 3: Paired Scatter plots Respected to Month