STAT 652: Predicting Flight Delays Project

Vincent Chiu

11/26/2019

Contents

1 Introduction **1**

2 Data **1**

3 Methods: **2**
- 3.1 Data Preprocessing . . . 2
- 3.2 Exploratory Analysis . . . 2
- 3.3 Principal Component Analysis (PCA) . . . 3
- 3.4 Validation . . . 3
- 3.5 Models . . . 3

4 Results **4**
- 4.1 GBM . . . 4
- 4.2 Linear Regression . . . 4
- 4.3 Comparison . . . 5

5 Conclusion and Discussion **5**
- 5.1 Discussion . . . 5
- 5.2 Conclusion . . . 5
- 5.3 Future Work . . . 5

6 Code **6**
- 6.1 Preparing the programming environment . . . 6
- 6.2 Data Preprocessing . . . 6
- 6.3 Feature Scaling . . . 8
- 6.4 Exploratory Data Analysis . . . 8
- 6.5 predicting 0 . . . 11
- 6.6 predicting the mean . . . 11
- 6.7 predicting the median . . . 11
- 6.8 linear regression with dest . . . 11
- 6.9 GBM . . . 12

References **17**

1 Introduction

The goal of this project is to predict the response variable, departure delays for a particular flight given the explanatory variables.

2 Data

The dataset consists of information about all the flights leaving from New York City in 2013. The dataset contains 43 variables in total. The dataset is an amalgamation of several datasets including datasets containing information on weather, the airports, the flights, and the models of airplanes. The training dataset provided to us contains 200,000 observations. Please see table 1 for what the data looks like.

3 Methods:

We will now outline the various methods used to clean and perform prediction on the data. We will discuss our techniques for data preprocessing, and cross validation, and the different models that we tried.

3.1 Data Preprocessing

I performed data preprocessing on the nycflights13 dataset. My data preprocessing steps include the following:

- loading the data from a csv
- setting the random seed for reproducibility of results
- casting all the columns with the character data type into the factor data type
- converting the shed_arr_time and sched_dep_time columns into the POSIX time format so that I can accurately take the difference of them.
- Dropping columns that contain data from after the planes' departure which may leak information about the response variable dep_delay. We drop columns "dep_time", "arr_time", "air_time", and "arr_delay".
- We drop column "year.x" because all the values are 2013
- We also drop tailnum because it produces too many dummy variable columns for one hot encoding.
- Dropping columns which consists of over 50% NAs which include the speed column. However, it should be noted that a rule of thumb suggested by Professor McNeney is to drop any columns with over 5% NAs. We use a different threshold for dropping columns leading to us keeping columns such as model instead of dropping it.
- Afterwards, I impute the missing values. However, there are limitations to this approach of imputing the missing values. It is possible, that the missingness of the plane model variable is related to dep_delay. In this scenario, we may be creating an inferior feature set by keeping the variable 'model' and imputing it. For example, say a highly unreliable plane model that frequently causes long delays has a high probability of being labeled as NA and represents the majority of NAs in the dataset. We would not be able to capture the relationship between this plane model and dep_delay if we imputed the model with the mode. To counteract this affect, we imputed NAs for the remaining columns using the imputeMissings library, adding a Boolean flag which indicates 1 if the associated value was 1 and 0 otherwise. For example, the "model_flag" for a given row is 1 if the "model" value was NA for that given row. Hence, no information is lost from our imputation.
- Normalizing the data to work well with methods like lasso regression.
- Only kept data which had a departure delay of less than 30 minutes late, which reduced the dataset from 200,000 rows to approximately 170,000. This is because we consider extreme delays of of over 30 minutes late to be freak accidents which cannot be accurately predicted by the available explanatory variables.

3.2 Exploratory Analysis

3.2.1 Correlations

First, we created a correlation plot for the numeric variables to see if there any correlations between the variables.

We see that there is very little correlation between the response variable dep_delay and any of the other variables. Some of the strongest correlations include the correlation between distance and longitude and time zone and a smaller correlation between distance and latitude. This makes sense as most of the planes are inter US flights from west to east or vice versa, there is not as much distance flown in the north south direction. Please see Figure 1 for the correlation plot (Click on the number after Figure to jump to plot).

3.3 Principal Component Analysis (PCA)

Next we performed PCA on only the numeric variables as techniques to perform PCA on mixed datasets (numerical and categorical) was not covered in class. When looking at the contribution of each variable to the first principal component, we notice that the variables lon, distance, tz, seats, alt, sched_air_time have the greatest absolute coefficients for the first principal component. The fact that the aforementioned variables have large coefficients in the first principal component suggests that they are highly correlated with each other. The fact that dep_delay has a small coefficient in the first principal component suggests that dep_delay is not highly correlated with any of the above variables.

As expected, it turns out that variables like lon, distance and tz are not important for predicting dep_delay according to the gbm model. This maybe be because although variables like lon, distance and tz help explain most of the variance in the dataset, they have a weak relationship with dep_delay.

Please see table 2 for the proportion of variance explained by each principal component. Please see table 3 for the coefficient of each variable for each principal component ordered by magnitude of coefficient.

3.4 Validation

Initially, I used the most basic validation technique where I have a training dataset and a cross validation dataset. I split the original data into a ratio of 2/3 train and 1/3 of the data for cross validation. There is a additional data which would be provided by the professor at a later date which we will use as the holdout test set. I believe that 2/3 of the data gives enough data for the models to train on while 1/3 is enough data for us to get an accurate assessment of the error. k-folds cross validation was not initially used in order to save on compute time as we were initially only exploring the models. k-folds cross validation would increase training time for the models by a factor of k. However, k-folds cross validation would lead to a more stable estimate of holdout test set error.

3.5 Models

We first explored some basic models to establish a baseline performance and compared it to our most sophisticated model, the Generalized Boosted Regression Model (GBM).

3.5.1 Basic Models

dep_delay is the number of minutes that the plane either departs early or late. Negative numbers are for early departures and positive numbers are for the number of minutes the plane is late. First, I used a basic model of simply predicting the dep_delay to always be 0. This was done to establish baseline performance. This model had an root mean squared error (RMSE) of 8.30571. TODO The model in which I predicted the mean for all the predictions had an RMSE of TODO.

3.5.2 Linear Regression

A linear regression model assumes a linear relationship between the explanatory variables and the response variable. The model minimizes the squared loss function. This minimization process generates coefficients which are used for a linear combintation of the explanatory variables. The linear combination is the prediction. Then I tried linear regression with dep_delay as the response variables and all the other remaining variables as the explanatory variables. This model was better than predicting the mean with an RMSE of TODO. This suggests that there is some relationship between the dep_delay and the explanatory variables.

3.5.3 Generalized Boosted Regression Model (GBM)

How boosted regression models work is we first start off with a baseline prediction. For example, we can use the mean as our baseline prediction. Then we use a base classifier to iteratively predict on the residuals multiplied by the shrinkage hyperparameter. Then this base classifier is added to the enssemble of base classifiers trained so far. A lower shrinkage effectively means a lower learning rate and therefore you need more iterations to reduce the train set residuals by the same amount. The benefit of a smaller shrinkage (with sufficient trees) is that you end up with a larger ensemble of trees that can reach a lower cross validation loss. This iterative prediction process is called boosting. In our case, our base classifiers are regressor trees. Each tree decides on its splitting criterion greedily by picking the split which results in the lowest mean squared error or some other splitting heuristic. We continue for n number of trees where n is specified by the user. Each iteration produces one tree, so the number of iterations is equal to the number of trees. Afterwards, we tried a Generalized Boosted Regression Model (GBM). This model had the lowest RMSE of 7.94458 on the cross validation set after it was tuned to have a shrinkage of 0.01 and around 16,000 trees. Shrinkage is proportional to the learning rate. 16,000 trees is the number of trees used in the model. Each iteration uses 1 tree, so 16,000 trees also refers to the number of iterations. According to the vignette, the RMSE can always be improved by decreasing shrinkage, but this provides diminishing returns. A good strategy would be to pick a small shrinkage that balances performance and compute time. Then with this fixed shrinkage value, increase the number of trees until you get diminishing returns. We decided to follow the aforementioned strategy. See figure 2 for a summary of our tuning experiments. Please see table 4 for a table of the relative influence of each explanatory variable. Here, you can see the relative influence for each variable for gbm.

For a gbm, the improvement in the splitting criterion (which is mean squared error for regression) for a given variable is calculated at each step. The relative influence for a given variable is the average of these improvements over all the trees where the aforementioned variable is used.

4 Results

In regression and gbm, I found different features to be important.
Please see table 5 for all the models and their root mean squared error (RMSE) on the validation set.

4.1 GBM

For the best gbm model, dest which refers to which airport a given plane was flying to was the most important feature. However, the one hot encoding versions of carrier were the most important features for regression. Based on the relative influence scores provided by the gbm, some of the most important feature variables include dest, model, and sched_dep_time_num_minute. The dest column contains the airport code for where a given flight is flying to. Based on my run of gbm with a shrinkage of 0.01 and 16834 trees, dest was the most important feature with 49.56 relative influence. ("Gradient Boosting Machines · UC Business Analytics R Programming Guide" 2019). dest is the destination airport code. 'sched_dep_time_num_minute' is the number of minutes since the beginning of a given day for that flight. 'model' is the plane nodel.

4.2 Linear Regression

On the other hand, dest does appear as an important feature in linear regression as well but it is not the most important feature. I surmise that if we can somehow sum up all the contributions from each of the one-hot-encoded variables derived from dest then, it might appear as the most important feature for linear regression as well. We can try using ANOVA in order to measure the statistical significance of dest. Performing ANOVA on comparing linear regression model with and without dest, it was determined that due to the low p-value of 0.0001863 associated with having dest that keeping at least one of the one hot categorical variables derived from dest is beneficial for the linear regression model.

4.3 Comparison

Dest was most important feature in gbm. It is possible that dest is important in combination with other variables which is something that the linear model without interaction terms cannot capture the relationship of whereas gbm can discover these non linear relationships.

TODO: try interaction terms , try anova.

5 Conclusion and Discussion

5.1 Discussion

We considered removing outliers in terms of 'dep_delay' in train but not in test, then use k-folds cross validation on test to determine how many outliers we should remove to boost performance on the cross-validation set. We considered removing highly influential points in order to train a better model. In this case, we consider highly influential points to be points with high cook's distances. However, this was infeasible as we did not have enough computational resources available and it took too long.

5.2 Conclusion

None of the models that we tried performed particularly well. We surmise that this may be due to the explanatory variables having a weak relationship with the 'dep_delay' variable. There is a lack of information about a particular flight before it reaches NYC. Instead, we get information about where the flight is going next which through commons sense, would reveal less information about the current condition of the plane and what kind of maintenance it would need and therefore what dep_delay it would have. Out of the methods that we covered in class, I found gradient boosted models to provide the best performance based on having the lowest root mean squared error on the cross validations et. I believe that this makes sense because GBMs are able to capture non linear relationships between the explanatory variables and 'dep_delay' whereas linear regression cannot.

5.3 Future Work

TODo: add dep_delay vs explanatory plots. TODO: remove points that are outliers ie dep_delay > 200 or 300 etc. or remove less than x number of points. then use k-folds cross validation on cross validation set where no points were removed. can repeat k-folds for different seeds. can just try this on my quickest model, i.e. linear regression. should be bowl shape vs RMSE vs. number of points removed. theoretically

I also considered removing based on cook's distance but this took too long to compute.

5 folds with 10 different random seeds

have train, CV and test set 1/3 train, 1/3 CV, 1/3 test 2/3% train, 1/3%CV, wait for prof test set

try lasso regression add more insights into generating dataset.

Conclusion stuff. certain planes tend to fly back and forth and that is why dest is good? interaction terms may be important because they are hard to predict in advance when you schedule flights which leads to delays? add more about shortcomings of approach. cut off too much of tail, reduce predictive power on test dataset which has tail in exchange for better performance on stuff not in tail.

maybe for stretch goal try rank stuff that prof did.

6 Code

6.1 Preparing the programming environment

6.1.1 Loading Libraries

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

6.2 Data Preprocessing

6.2.1 Loading the data

```
library(nycflights13)
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
set.seed(42)
original_data <- read_csv("fltrain.csv.gz")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   carrier = col_character(),
```

Table 1: A table of the first few rows of the nycflights13 data.

year.x	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay	carrier
2013	11	7	600	600	0	826	825	1	WN
2013	10	30	1252	1250	2	1356	1400	-4	AA
2013	12	18	1723	1715	8	2008	2020	-12	DL
2013	11	20	2029	2030	-1	2141	2205	-24	WN
2013	10	21	1620	1625	-5	1818	1831	-13	DL
2013	11	7	852	900	-8	1139	1157	-18	B6

```
##   tailnum = col_character(),
##   origin = col_character(),
##   dest = col_character(),
##   time_hour = col_datetime(format = ""),
##   name = col_character(),
##   dst = col_character(),
##   tzone = col_character(),
##   type = col_character(),
##   manufacturer = col_character(),
##   model = col_character(),
##   engine = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
DF <- original_data
```

turning all columns with datatype characters to factors.

```
DF[sapply(DF, is.character)] <- lapply(DF[sapply(DF, is.character)],
                                       as.factor)
DF$flight <- as.factor(DF$flight)
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
DF$sched_arr_time_posix <- as.POSIXct(str_pad(as.character(DF$sched_arr_time), 4, pad="0"),format="%H%M
DF$sched_arr_time_hour <- hour(DF$sched_arr_time_posix)
DF$sched_arr_time_minute <- minute(DF$sched_arr_time_posix)

#num minute is number of minutes since start of day for scheduled arrival time
DF$sched_arr_time_num_minute <- 60*DF$sched_arr_time_hour + DF$sched_arr_time_minute

DF$sched_dep_time_posix <- as.POSIXct(str_pad(as.character(DF$sched_dep_time),4 , pad="0"),format="%H%M
DF$sched_dep_time_hour <- hour(DF$sched_dep_time_posix)
DF$sched_dep_time_minute <- minute(DF$sched_dep_time_posix)
#num minute is number of minutes since start of day for scheduled depival time
DF$sched_dep_time_num_minute <- 60*DF$sched_dep_time_hour + DF$sched_dep_time_minute
```

```r
select(original_data, time_hour, sched_dep_time, sched_arr_time, tz, tzone)
select(DF, sched_arr_time, sched_arr_time_hour)
```

```r
DF$sched_air_time <- DF$sched_arr_time_posix - DF$sched_dep_time_posix
drops <- c('sched_arr_time_posix', 'sched_arr_time_hour', 'sched_dep_time_posix', 'sched_dep_time_hour'
DF <- DF[ , !(names(DF) %in% drops)]
```

```r
drops <- c("dep_time", "arr_time", "air_time", "arr_delay", "year.x", 'tailnum')
DF <- DF[ , !(names(DF) %in% drops)]
```

```r
## Remove columns with more than 50% NA
DF <- DF[, -which(colMeans(is.na(DF)) > 0.5)]
```

```r
DF$sched_air_time <- as.numeric(DF$sched_air_time)
library(imputeMissings)
```

```
##
## Attaching package: 'imputeMissings'

## The following object is masked from 'package:Hmisc':
##
##     impute

## The following object is masked from 'package:dplyr':
##
##     compute
```

```r
impute_model <- imputeMissings::compute(DF, method="median/mode")
impute_model
DF <- impute(DF, object=impute_model, flag=TRUE)
DF <- DF[!duplicated(as.list(DF))]  #remove all redundant flag columns that are identical to each other
```

```r
numeric_only_df <- dplyr::select_if(DF, is.numeric)
library(corrplot)
```

```
## corrplot 0.84 loaded
```

6.3 Feature Scaling

```r
dep_delay_vec <- DF$dep_delay
DF$dep_delay <- NULL
head(DF)

library(dplyr)
DF <- DF %>% mutate_if(is.numeric, scale)
head(DF)
DF$dep_delay <- dep_delay_vec
```

6.4 Exploratory Data Analysis

```r
numeric_DF <- dplyr::select_if(DF, is.numeric) %>%  scale()
```

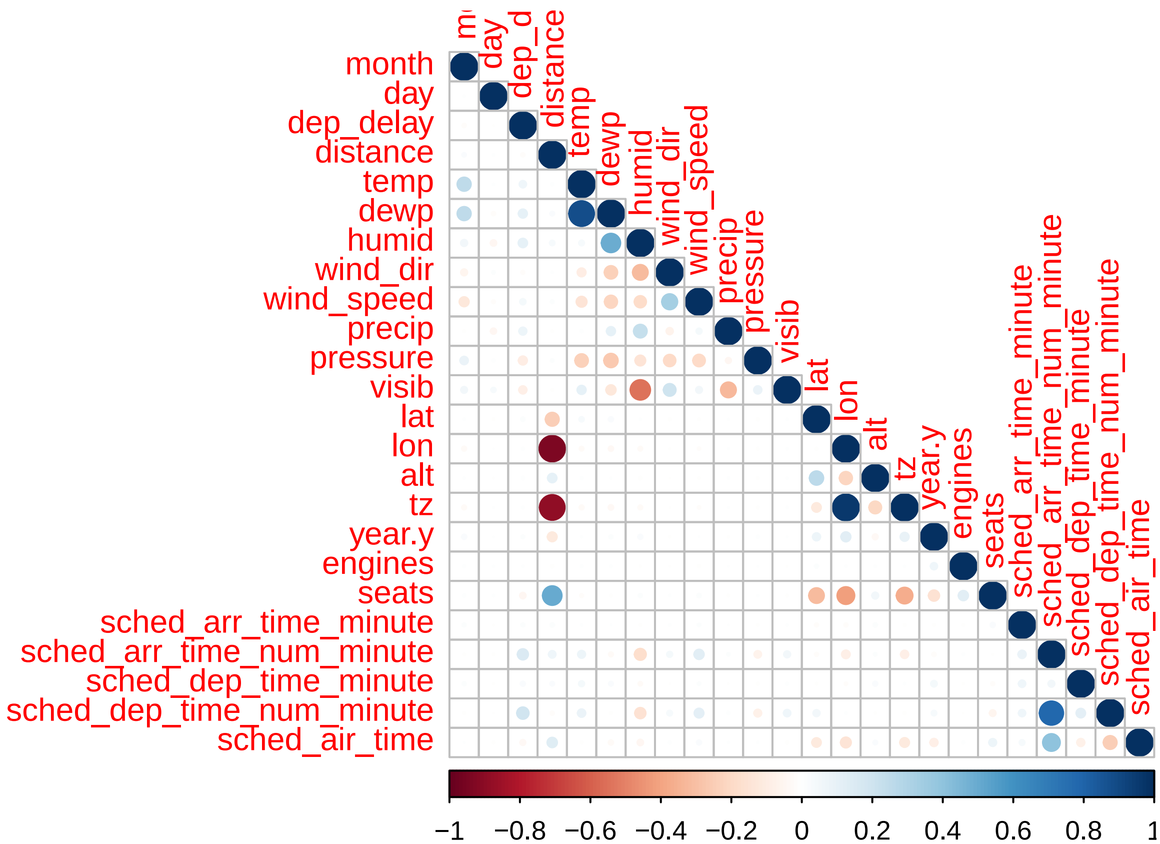


Figure 1: grid depicting correlation amongst all numerical variables

Table 2: proportion of variance explained by each principal component

	x
PC1	0.1352687
PC2	0.1061946
PC3	0.0862791
PC4	0.0688860
PC5	0.0605903
PC6	0.0582950

Table 3: coefficients for each variable on each principal component

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
dewp	-0.0345819	0.5398436	-0.2228860	0.1989663	-0.0782615	-0.1213865	0.0334310	0.0076414
humid	-0.0203752	0.4533838	0.1449535	-0.3156007	0.0037085	-0.0164143	-0.0179066	-0.0058038
temp	-0.0291751	0.3840823	-0.3360928	0.3735115	-0.0921201	-0.1323152	0.0490440	0.0106687
wind_dir	0.0031602	-0.2944783	-0.1128694	0.0491077	-0.0627922	-0.4851444	0.0127712	-0.0479663
visib	0.0071745	-0.2801986	-0.1460066	0.5175333	-0.0077024	-0.0187428	-0.0264190	0.0248309
wind_speed	-0.0082656	-0.2533491	-0.1460835	-0.2029687	-0.0619018	-0.4378020	0.0093614	-0.0472185

```
prcomp_res <- prcomp(numeric_DF)
sdev <- prcomp_res$sdev
sdev
```

6.4.1 all four components at same time

proportion of variance explained by each component

```
pve <- colSums(prcomp_res$x^2)/sum(numeric_DF^2)
```

```
rotation <- as.data.frame(prcomp_res$rotation)
rotation[order(-abs(rotation$PC1)),]
```

```
pca_rotation <- head(rotation[order(-abs(rotation$PC2)),])
```

6.4.2 take out extreme departure delays

```
DF<-DF[DF$dep_delay < 30,]
```

```
set.seed(42)
DF$flight <- NULL
train_index <- sample(1:nrow(DF),size=2*nrow(DF)/3,replace=FALSE)
train_df <- DF[train_index,]
test_df <- DF[-train_index,]
```

```
# pre-allocate space
preallocate_df <- function(n){
  df <- data.frame(model_description = character(n), rmse = numeric(n), stringsAsFactors = FALSE)
  for(i in 1:n){
```

```
    df$model_description[i] <- i
    df$rmse[i] <- toString(i)
  }
  df
}
```

6.5 predicting 0

```
benchmark_df <- data.frame(model_description = character(), rmse = numeric(), stringsAsFactors = FALSE)
rmse = mean((test_df$dep_delay-0)^2) %>% sqrt()
model_description = "predicting 0"
benchmark_df <- rbind(benchmark_df, data.frame(model_description = model_description, rmse=rmse))
```

6.6 predicting the mean

```
rmse = mean((test_df$dep_delay-mean(train_df$dep_delay))^2)%>% sqrt()
rmse
benchmark_df
```

6.7 predicting the median

```
rmse = mean((test_df$dep_delay-median(train_df$dep_delay))^2)%>% sqrt()
rmse
model_description <- 'predicting the median'
benchmark_df <- rbind(benchmark_df, data.frame(model_description = model_description, rmse=rmse))
```

6.8 linear regression with dest

```
model <- lm(dep_delay ~ ., data=train_df)
model_without_dest <-  lm(dep_delay ~ .-dest, data=train_df)
anova(model, model_without_dest)
summary <- round(summary(model)$coefficients,6)
sorteddf <- summary[order(summary[,ncol(summary)]),]
head(sorteddf)
```

```
lm_test_df <- test_df

in_test_but_not_train <- setdiff(unique(lm_test_df$model), unique(train_df$model))
lm_test_df <- lm_test_df[ !lm_test_df$model %in% in_test_but_not_train, ]

in_test_but_not_train <- setdiff(unique(lm_test_df$dest), unique(train_df$dest))
lm_test_df <- lm_test_df[ !lm_test_df$dest %in% in_test_but_not_train, ]

preds = predict(model, newdata=lm_test_df)
```

```
## Warning in predict.lm(model, newdata = lm_test_df): prediction from a rank-
## deficient fit may be misleading
```

```
rmse = sqrt(mean((lm_test_df$dep_delay - preds)^2))
rmse
model_description <- 'linear regression'
benchmark_df <- rbind(benchmark_df, data.frame(model_description = model_description, rmse=rmse))
```

6.9 GBM

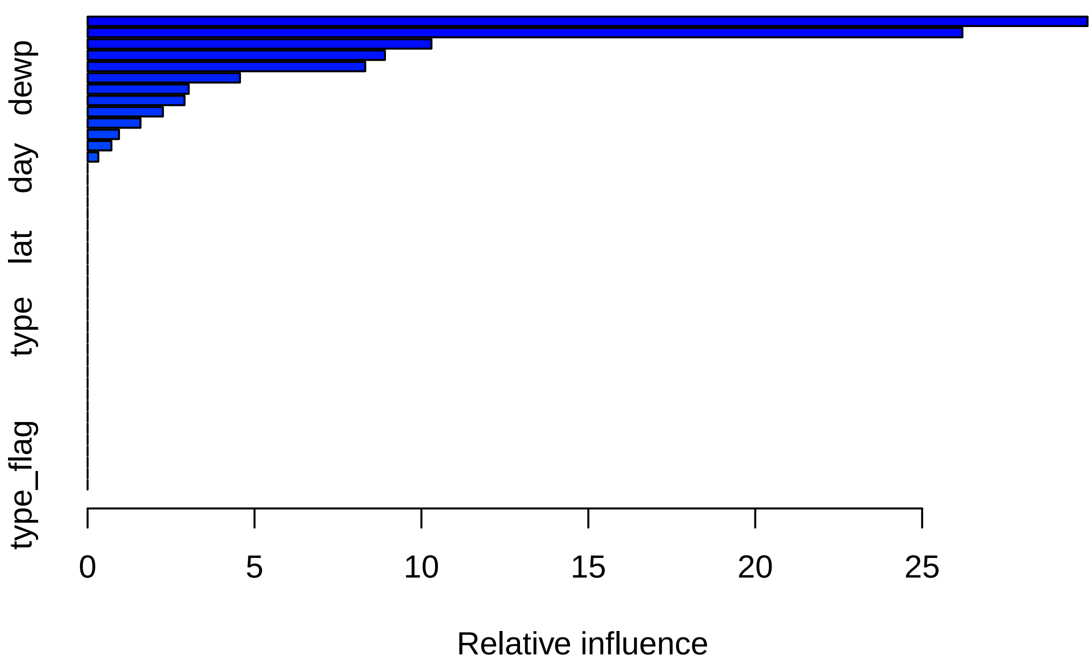
```
library(gbm)
```

```
## Loaded gbm 2.1.5
```

```
train_gbm <- function(filename){
set.seed(42)
model <- gbm(dep_delay ~ ., data=train_df,
              n.trees=100, shrinkage=0.1) # default shrinkage = 0.1
preds = predict(model, newdata=test_df, n.trees=100)
rmse = sqrt(mean((test_df$dep_delay - preds)^2))
summary(model)
saveRDS(model, filename)
return(model)
}

destfile <- "models/model2.rds"
if (!file.exists(destfile)) {
   train_gbm(destfile)
 }
model <- readRDS(destfile)
```

```
benchmark_df <- rbind(benchmark_df, data.frame(model_description = 'gbm', rmse=7.94458))
```



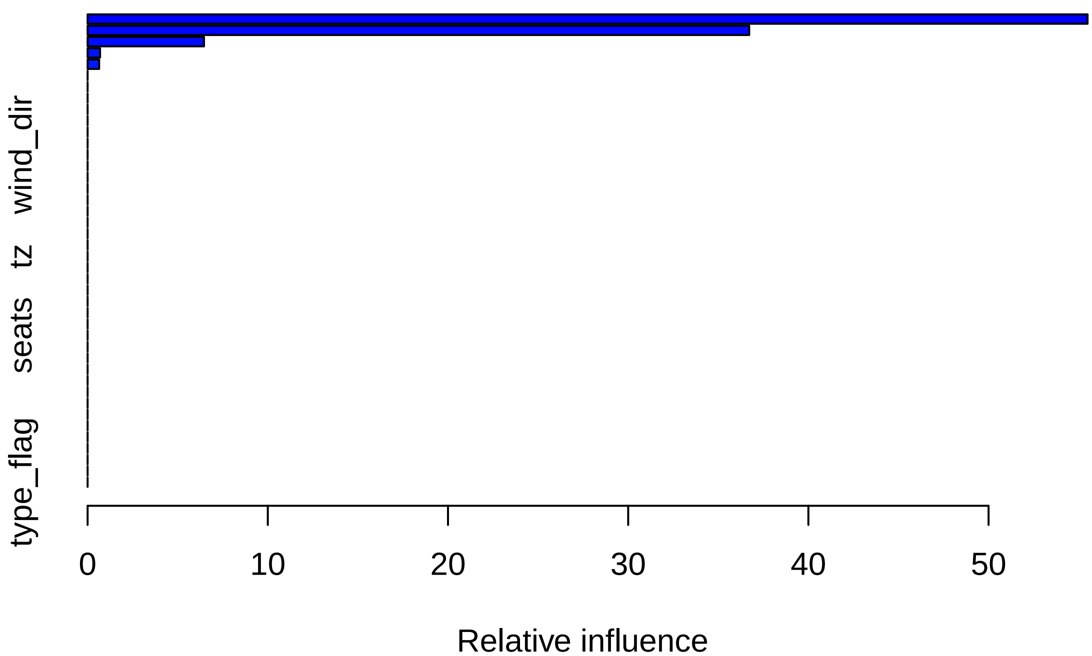
Here, you can see the relative influence for each variable for gbm.
For a gbm, the improvement in the splitting criterion (which is mean squared error for regression) for a

Table 4: gbm relative influence

	var	rel.inf
sched_dep_time_num_minute	sched_dep_time_num_minute	29.955850
model	model	26.206955
dest	dest	10.299706
carrier	carrier	8.907853
month	month	8.317570
dewp	dewp	4.563640

given variable is calculated at each step. The relative influence for a given variable is the average of these improvements over all the trees where the aforementioned variable is used.

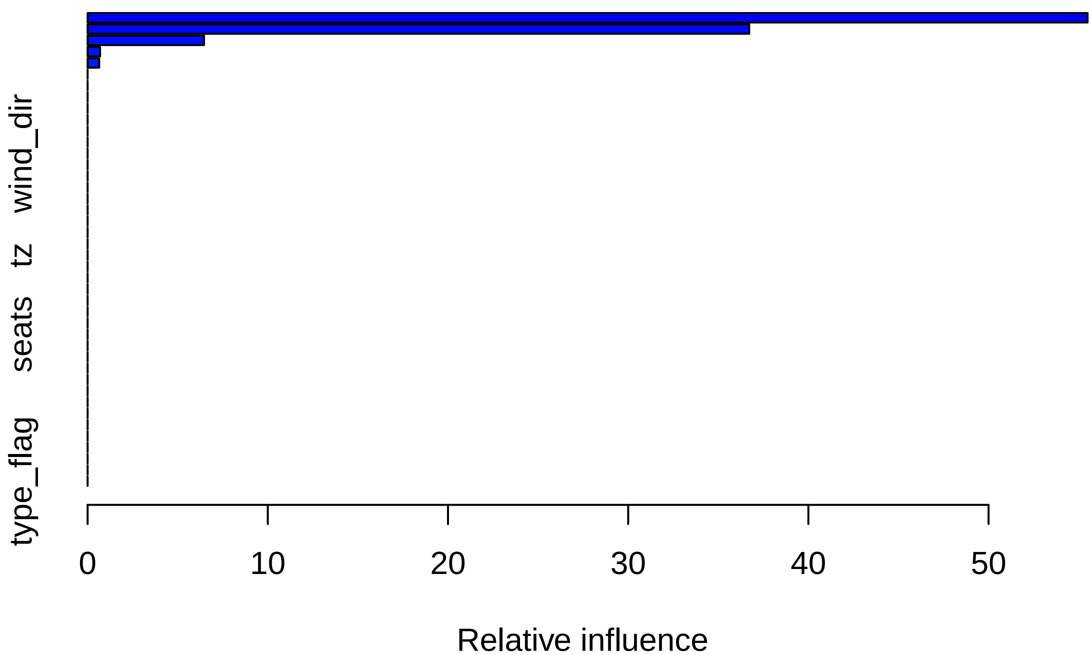
```
model <- gbm(dep_delay ~ ., data=train_df,
             n.trees=100, shrinkage=0.01) # default shrinkage = 0.1
preds = predict(model, newdata=test_df, n.trees=100)
filename <- "models/gbm_shrinkage_0point01_ntrees_100_v1.rds"
saveRDS(model, filename)
rmse = sqrt(mean((test_df$dep_delay - preds)^2))
summary(model)
```




```
library(gbm)
filename <- "models/gbm_shrinkage_0point01_ntrees_100_v1.rds"
if (!file.exists(filename)) {
  model <- gbm(
    dep_delay ~ .,
    data = train_df,
    n.trees = 100,
    shrinkage = 0.01
  ) # default shrinkage = 0.1
  saveRDS(model, filename)
```

```r
} else {
  print("reading saved model")
  model <- readRDS(filename)
}
preds <- predict(model, newdata = test_df, n.trees = 100)
rmse <- sqrt(mean((test_df$dep_delay - preds) ^ 2))

summary(model)
```




```r
rmse = sqrt(mean((test_df$dep_delay - preds)^2))
```

```r
cars %>%
  rename("Stopping Distance (ft)" = dist) %>%
  colnames()
```

```r
shrinkage_0point01_bench <- read.csv('performance/shrinkage_0point01_numtrees_32_to_16384_gbm_benchmark
shrinkage_0point01_bench$X <- NULL
shrinkage_0point01_bench <- shrinkage_0point01_bench %>%
  dplyr::rename(
    "shrinkage_0point01_rmse" = rmse
    )
shrinkage_0point01_bench
```

```r
shrinkage_0point001_bench <- read.csv('performance/shrinkage_0point001_numtrees_32_to_16384_gbm_benchma
shrinkage_0point001_bench$X <- NULL
shrinkage_0point001_bench <- shrinkage_0point001_bench %>%
  dplyr::rename(
    "shrinkage_0point001_rmse" = rmse
    )
shrinkage_0point001_bench
```

```r
gbm_merge_df <- merge(shrinkage_0point01_bench, shrinkage_0point001_bench)
gbm_merge_df
```

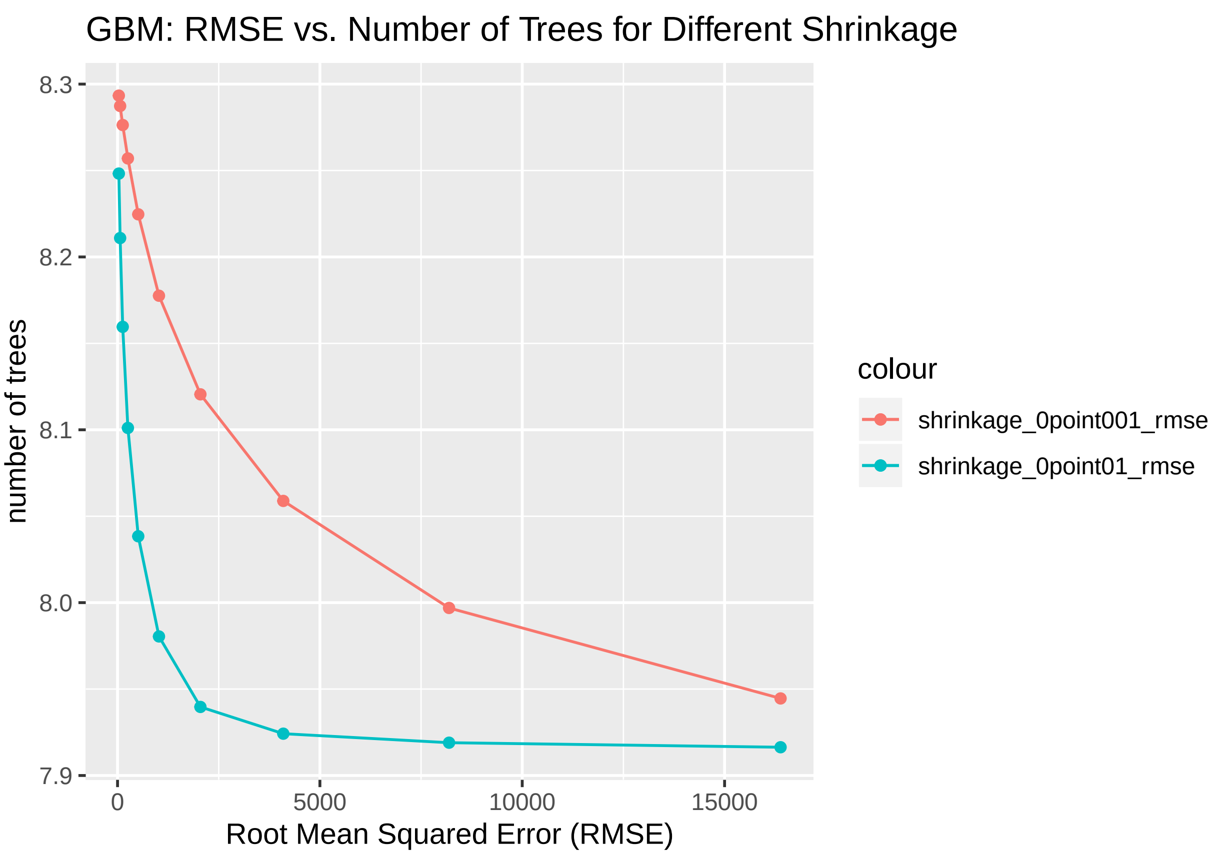


Figure 2: plot of RMSE vs number of trees for shrinkage = 0.01 and shrinkage = 0.001

Table 5: benchmark comparing all the models that we tried

model_description	rmse
predicting 0	8.305710
predicting the median	8.469257
linear regression	7.989994
gbm	7.944580

```r
library(gbm)
rerun <- TRUE
set.seed(42)
x <- 2 ^ seq(5, 14, by = 1)
rmse_vec <- numeric(length(x))
count <- 1
for (val in x) {
  filename <-
    paste0("models/gbm_shrinkage_0point001_ntrees_", val)
  filename <- paste0(filename, "_v2.rds")
  if (!file.exists(filename) | rerun) {
    hboost <- gbm(
      dep_delay ~ .,
      data = train_df,
      n.trees = val,
      distribution = 'gaussian',
      shrinkage = 0.001
    )
    saveRDS(hboost, filename)
  } else{
    hboost <- readRDS(filename)
  }

  preds = predict(hboost, n.trees = val, newdata = test_df)
  mse = mean((test_df$dep_delay - preds) ^ 2)
  rmse <- sqrt(mse)
  rmse_vec[count] <- rmse
  print(val)
  print(rmse)
  count = count + 1
}

summary <- summary(hboost)
write.csv(summary,
          'performance/gbm_shrinkage_0point001_16384trees_summary_v1.csv')

plot(x, rmse_vec)
num_trees_vs_rmse <- data.frame("num_trees" = x, "rmse" = rmse_vec)
write.csv(
  num_trees_vs_rmse,
  'performance/gbm_shrinkage_0point001_num_trees_vs_rmse_v1.csv'
)
```

References

"Gradient Boosting Machines · UC Business Analytics R Programming Guide." 2019. http://uc-r.github.io/gbm_regression#h2o.