# Combined Model by a Meta-Regressor

*Hannah Khuong*

## Model Summary

I used a meta-linear-regressor to stack the results from the selected models (sub-learners). For each Days Prior period, the meta-regression was run to combine different prediction patterns of these individual models. For example, for New York hotel, meta-regressors was run on 5 predictions provided by 5 models for data in Day 01-07 period, Day 08-14 period and so on. The predictions of all Days Prior periods were then combined to be the prediction for all 60 days prior to the Stay Date.

A strength of this combination method is the regression can combine pattern that one model can capture with a pattern another model captures that the first model failed to capture.

The 5 models (sub-learners) used in this meta-regression are Traditional, Regression, KNN, Tree, and Random Forest. Our stacked model is a successful one because this model outperformed all the sub-learners' predictions in all Days Prior periods.

I obtained the best validation results from this model. In every Days Prior period, the out-sample MASEs were the smallest. This result indicates a much better predictive ability of this combined model compared to Naïve prediction. The best Days Prior period is Day 01-07, which sucessfully recued 80% error in Naïve predictions.

At the end of this file are visualization of actual and predicted cancellation rate produced by this Combined model.

## Load Predictions from other Models

```
train_reg_pred <- read.csv("../../treated data/train_regre_pred.csv", header=TRUE, sep=",")
train_reg_pred <- train_reg_pred[2]
train_tree_pred <- read.csv("../../treated data/train_nyc_tree_pred.csv", header=TRUE, sep=",")
train_tree_pred <- train_tree_pred[2]
train_trad_pred <- read.csv("../../treated data/train_trad_pred.csv", header=TRUE, sep=",")
train_trad_pred <- train_trad_pred[2]
train_rf_pred <- read.csv("../../treated data/train_rf_pred.csv", header=TRUE, sep=",")
train_rf_pred <- train_rf_pred[2]
train_nn_pred <- read.csv("../../treated data/train_nn_pred.csv", header=TRUE, sep=",")
train_nn_pred <- train_nn_pred[2]
train_nyc <- read.csv("../../treated data/train_nyc.csv", header=TRUE, sep=",")
train_knn_pred <- train_nyc['knn_pred']
train_true_cxl <- train_nyc['cxl_rate']
train_days_prior <- train_nyc['days_prior']

test_reg_pred <- read.csv("../../treated data/test_regre_pred.csv", header=TRUE, sep=",")
test_reg_pred <- test_reg_pred[2]
test_tree_pred <- read.csv("../../treated data/test_nyc_tree_pred.csv", header=TRUE, sep=",")
test_tree_pred <- test_tree_pred[2]
test_trad_pred <- read.csv("../../treated data/test_trad_pred.csv", header=TRUE, sep=",")
test_trad_pred <- test_trad_pred[2]
test_rf_pred <- read.csv("../../treated data/test_rf_pred.csv", header=TRUE, sep=",")
test_rf_pred <- test_rf_pred[2]
```

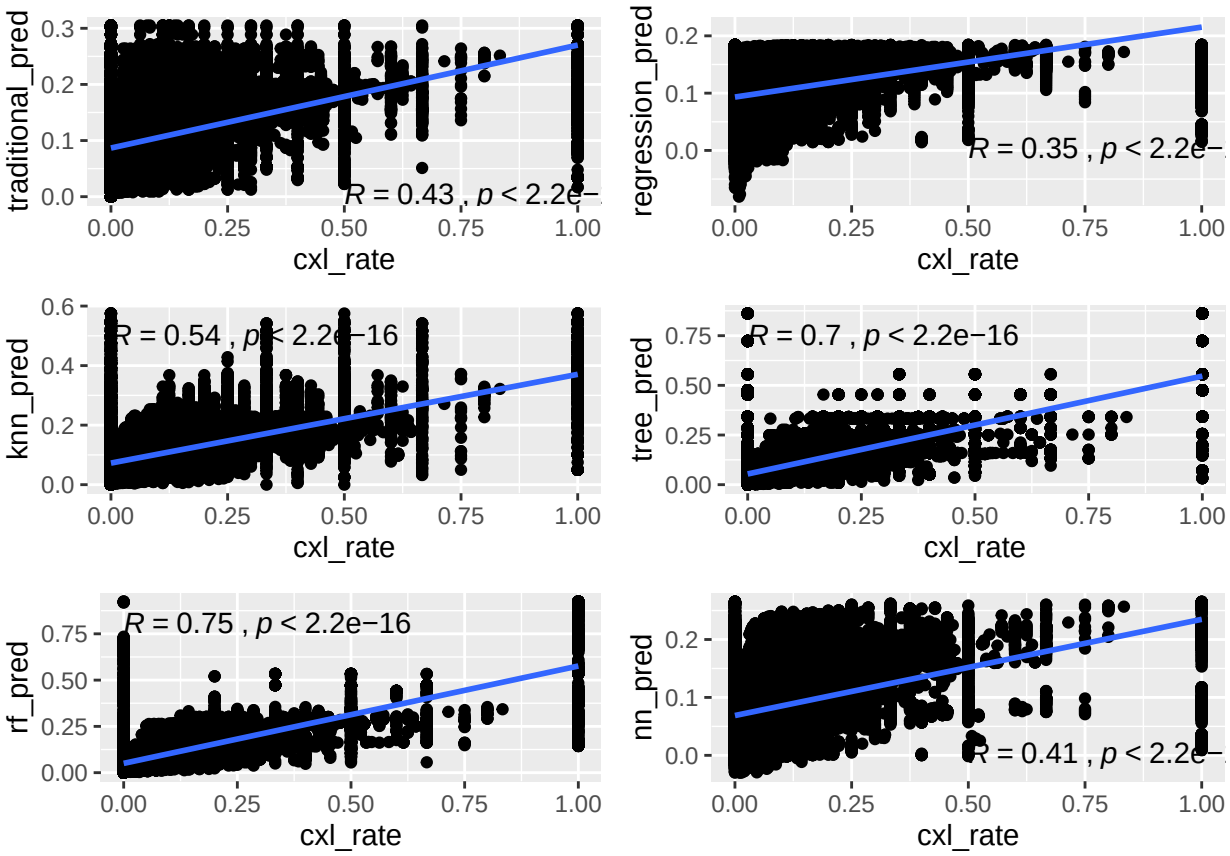
```
test_nn_pred <- read.csv("../../treated data/test_nn_pred.csv", header=TRUE, sep=",")
test_nn_pred <- test_nn_pred[2]
test_nyc <- read.csv("../../treated data/test_nyc.csv", header=TRUE, sep=",")
test_knn_pred <- test_nyc['knn_pred']
test_true_cxl <- test_nyc['cxl_rate']
test_days_prior <- test_nyc['days_prior']


train_nyc <- train_nyc %>% select(-c('X', 'knn_pred'))
test_nyc <- test_nyc %>% select(-c('X', 'knn_pred'))
```

## Combine Predictions into a data set

```
# create set with all predictions
train_pred_set <- cbind(train_nyc, train_reg_pred, train_tree_pred, train_trad_pred,
                        train_knn_pred, train_rf_pred, train_nn_pred)

test_pred_set <- cbind(test_nyc, test_reg_pred, test_tree_pred, test_trad_pred,
                       test_knn_pred,  test_rf_pred, test_nn_pred)
```

### Visualize predictions

```
# Visualize prediction against actual cxl rate
grid.arrange(
train_pred_set %>% ggplot(aes(x = cxl_rate, y = traditional_pred))+geom_point()+
  geom_smooth(method = "lm")+ stat_cor(method = "pearson", label.x = 0.5, label.y = 0),

train_pred_set %>% ggplot(aes(x = cxl_rate, y = regression_pred))+geom_point()+
  geom_smooth(method = "lm")+ stat_cor(method = "pearson", label.x = 0.5, label.y = 0),

train_pred_set %>% ggplot(aes(x = cxl_rate, y = knn_pred))+geom_point()+
  geom_smooth(method = "lm")+ stat_cor(method = "pearson"),

train_pred_set %>% ggplot(aes(x = cxl_rate, y = tree_pred))+geom_point()+
  geom_smooth(method = "lm")+ stat_cor(method = "pearson"),

train_pred_set %>% ggplot(aes(x = cxl_rate, y = rf_pred))+geom_point()+
  geom_smooth(method = "lm")+ stat_cor(method = "pearson"),

train_pred_set %>% ggplot(aes(x = cxl_rate, y = nn_pred))+geom_point()+
  geom_smooth(method = "lm")+ stat_cor(method = "pearson", label.x = 0.5, label.y = 0), ncol = 2)
```




## Meta-Regressor for each days prior period

### Day 1-7

```
# Create set for selected days
train_day_1_7 <- train_pred_set %>% filter(days_prior <= 7) %>%
  select(traditional_pred, tree_pred, cxl_rate, knn_pred, rf_pred, stay_dt, product_type, days_prior,
         regression_pred, OTB, naive_survive_pred, OTB_to_survive)
test_day_1_7 <- test_pred_set %>% filter(days_prior <= 7) %>%
  select(traditional_pred,  tree_pred,cxl_rate, knn_pred, rf_pred, stay_dt, product_type,
         days_prior,regression_pred, OTB, naive_survive_pred, OTB_to_survive)

# Regression
day_1_7_reg <- lm(cxl_rate ~ traditional_pred + tree_pred +rf_pred+regression_pred, data=train_day_1_7)
day_1_7_reg %>% summary()
```

```
##
## Call:
## lm(formula = cxl_rate ~ traditional_pred + tree_pred + rf_pred +
##     regression_pred, data = train_day_1_7)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38049 -0.01994 -0.00253  0.01229  0.87193
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.007147   0.001919  -3.724 0.000198 ***
## traditional_pred  0.154808   0.052658   2.940 0.003298 **
## tree_pred        -0.137280   0.034120  -4.024 5.82e-05 ***
## rf_pred           1.075688   0.043875  24.517  < 2e-16 ***
## regression_pred  -0.037683   0.053926  -0.699 0.484717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07736 on 5119 degrees of freedom
## Multiple R-squared:  0.3103, Adjusted R-squared:  0.3097
## F-statistic: 575.7 on 4 and 5119 DF,  p-value: < 2.2e-16
```

```
# Predict and assign to new df
train_day_1_7$combine_pred <- predict(day_1_7_reg, train_day_1_7)
test_day_1_7$combine_pred <- predict(day_1_7_reg, test_day_1_7)

train_day_1_7 <- train_day_1_7 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as

test_day_1_7 <- test_day_1_7 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as
```

### Day 8-14

```
# Create set for selected days
train_day_8_14 <- train_pred_set %>% filter(days_prior > 7 & days_prior <= 14) %>%
  select(traditional_pred, tree_pred, cxl_rate,rf_pred, knn_pred, stay_dt, product_type,
         days_prior,regression_pred, OTB, naive_survive_pred, OTB_to_survive)
test_day_8_14 <- test_pred_set %>% filter(days_prior > 7 & days_prior <= 14) %>%
  select(traditional_pred, tree_pred, knn_pred, rf_pred, stay_dt, product_type, days_prior,regression_pr

# Regression
day_8_14_reg <- lm(cxl_rate ~ traditional_pred  +knn_pred+rf_pred+regression_pred, data=train_day_8_14)
day_8_14_reg %>% summary()
```

```
##
## Call:
## lm(formula = cxl_rate ~ traditional_pred + knn_pred + rf_pred +
##     regression_pred, data = train_day_8_14)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.51150 -0.02910 -0.00201  0.02172  0.82829
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.003674   0.003771   0.974   0.3299
## traditional_pred -0.320021   0.071748  -4.460 8.36e-06 ***
## knn_pred          0.212918   0.042540   5.005 5.77e-07 ***
## rf_pred           1.199800   0.028489  42.114  < 2e-16 ***
```

```
## regression_pred  -0.144668   0.083649  -1.729   0.0838 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09969 on 5119 degrees of freedom
## Multiple R-squared:  0.3604, Adjusted R-squared:  0.3599
## F-statistic:   721 on 4 and 5119 DF,  p-value: < 2.2e-16
```

```
# Predict and assign to df
train_day_8_14$combine_pred <- predict(day_8_14_reg, train_day_8_14)
test_day_8_14$combine_pred <- predict(day_8_14_reg, test_day_8_14)

train_day_8_14 <- train_day_8_14 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as

test_day_8_14 <- test_day_8_14 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as
```

### Day 15-20

```
# Create set for selected days
train_day_15_20 <- train_pred_set %>% filter(days_prior > 14 & days_prior <= 20) %>%
  select(traditional_pred, tree_pred, cxl_rate,rf_pred, knn_pred, stay_dt, product_type,
         days_prior,regression_pred, OTB, naive_survive_pred, OTB_to_survive)
test_day_15_20 <- test_pred_set %>% filter(days_prior > 14 & days_prior <= 20) %>%
  select(traditional_pred, tree_pred, knn_pred, rf_pred, stay_dt, product_type, days_prior,
         cxl_rate,regression_pred, OTB, naive_survive_pred, OTB_to_survive)

# Regression
day_15_20_reg <- lm(cxl_rate ~ traditional_pred  +knn_pred+rf_pred+regression_pred, data=train_day_15_20
day_15_20_reg %>% summary()
```

```
##
## Call:
## lm(formula = cxl_rate ~ traditional_pred + knn_pred + rf_pred +
##     regression_pred, data = train_day_15_20)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70504 -0.03377 -0.00019  0.02691  0.81913
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.003721   0.004046   0.920  0.35779
## traditional_pred -0.167878   0.063648  -2.638  0.00838 **
## knn_pred         -0.106838   0.040697  -2.625  0.00869 **
## rf_pred           1.351310   0.027592  48.976  < 2e-16 ***
## regression_pred  -0.112916   0.079135  -1.427  0.15369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09538 on 4387 degrees of freedom
## Multiple R-squared:  0.4858, Adjusted R-squared:  0.4853
```

```
## F-statistic:  1036 on 4 and 4387 DF,  p-value: < 2.2e-16
```

```
# Predict and assign to df
train_day_15_20$combine_pred <- predict(day_15_20_reg, train_day_15_20)
test_day_15_20$combine_pred <- predict(day_15_20_reg, test_day_15_20)

train_day_15_20 <- train_day_15_20 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as

test_day_15_20 <- test_day_15_20 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as
```

### Day 21-27

```
# Create set for selected days
train_day_21_27 <- train_pred_set %>% filter(days_prior > 20 & days_prior <= 27) %>%
  select(traditional_pred, tree_pred, cxl_rate,rf_pred, knn_pred, stay_dt, product_type,
         days_prior, regression_pred,OTB, naive_survive_pred, OTB_to_survive)
test_day_21_27 <- test_pred_set %>% filter(days_prior > 20 & days_prior <= 27) %>%
  select(traditional_pred, tree_pred, knn_pred, rf_pred, stay_dt, product_type,
         cxl_rate,days_prior,regression_pred, OTB, naive_survive_pred, OTB_to_survive)

# Regression
day_21_27_reg <- lm(cxl_rate ~ traditional_pred  +knn_pred+rf_pred+regression_pred, data=train_day_21_27
day_21_27_reg %>% summary()
```

```
##
## Call:
## lm(formula = cxl_rate ~ traditional_pred + knn_pred + rf_pred +
##     regression_pred, data = train_day_21_27)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63873 -0.03685  0.00100  0.02692  0.84421
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.004460   0.003846   1.160   0.2463
## traditional_pred  0.190127   0.043151   4.406 1.07e-05 ***
## knn_pred         -0.089147   0.035576  -2.506   0.0122 *
## rf_pred           1.132143   0.022737  49.792  < 2e-16 ***
## regression_pred  -0.290335   0.055193  -5.260 1.50e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1009 on 5119 degrees of freedom
## Multiple R-squared:  0.4742, Adjusted R-squared:  0.4737
## F-statistic:  1154 on 4 and 5119 DF,  p-value: < 2.2e-16
```

```
# Predict and assign to df
train_day_21_27$combine_pred <- predict(day_21_27_reg, train_day_21_27)
test_day_21_27$combine_pred <- predict(day_21_27_reg, test_day_21_27)

train_day_21_27 <- train_day_21_27 %>%
```

```
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as

test_day_21_27 <- test_day_21_27 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as
```

### Day 28-60

```
# Create set for selected days
train_day_28_60 <- train_pred_set %>% filter(days_prior > 27 ) %>%
  select(traditional_pred, tree_pred, cxl_rate,rf_pred, knn_pred, stay_dt, product_type,
         days_prior, regression_pred,OTB, naive_survive_pred, OTB_to_survive)
test_day_28_60 <- test_pred_set %>% filter(days_prior > 27 ) %>%
  select(traditional_pred, tree_pred, knn_pred, rf_pred, stay_dt, product_type, days_prior,
         cxl_rate, regression_pred,OTB, naive_survive_pred, OTB_to_survive)

# Regression
day_28_60_reg <- lm(cxl_rate ~ knn_pred  + rf_pred+tree_pred+regression_pred, data=train_day_28_60)
day_28_60_reg %>% summary()
```

```
##
## Call:
## lm(formula = cxl_rate ~ knn_pred + rf_pred + tree_pred + regression_pred,
##     data = train_day_28_60)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00864 -0.03632  0.00443  0.02106  0.94953
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.002958   0.001792  -1.651  0.09876 .
## knn_pred         0.021839   0.010990   1.987  0.04692 *
## rf_pred          1.237877   0.020452  60.527  < 2e-16 ***
## tree_pred       -0.177201   0.019921  -8.895  < 2e-16 ***
## regression_pred -0.045803   0.015339  -2.986  0.00283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1204 on 24151 degrees of freedom
## Multiple R-squared:  0.6042, Adjusted R-squared:  0.6041
## F-statistic:  9215 on 4 and 24151 DF,  p-value: < 2.2e-16
```

```
# Predict and assign to df
train_day_28_60$combine_pred <- predict(day_28_60_reg, train_day_28_60)
test_day_28_60$combine_pred <- predict(day_28_60_reg, test_day_28_60)

train_day_28_60 <- train_day_28_60 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as

test_day_28_60 <- test_day_28_60 %>%
  mutate(combine_pred = ifelse(combine_pred < 0, 0, combine_pred)) # Assign negative predicted value as
```

```r
train_combine <- rbind(train_day_1_7, train_day_8_14, train_day_15_20, train_day_21_27, train_day_28_60)
test_combine <- rbind(test_day_1_7, test_day_8_14, test_day_15_20, test_day_21_27, test_day_28_60)
```

## Metrics Summary

```r
train_combine <- train_combine %>%
   mutate(days_prior_cat = case_when(days_prior <= 7  ~ "Day 01-07",
                          days_prior > 7 & days_prior <= 14 ~ "Day 08-14",
                          days_prior > 14 & days_prior <= 20 ~ "Day 15-20",
                          days_prior > 20 & days_prior <= 27 ~ "Day 21-27",
                          days_prior > 27 & days_prior <= 60 ~ "Day 28-60"))

# Establish order for days_prior_cat
train_combine$days_prior_cat <- factor(train_combine$days_prior_cat,
                  levels = c('Day 01-07', 'Day 08-14', 'Day 15-20', 'Day 21-27', 'Day 28-60'))

test_combine <- test_combine %>%
   mutate(days_prior_cat = case_when(days_prior <= 7  ~ "Day 01-07",
                          days_prior > 7 & days_prior <= 14 ~ "Day 08-14",
                          days_prior > 14 & days_prior <= 20 ~ "Day 15-20",
                          days_prior > 20 & days_prior <= 27 ~ "Day 21-27",
                          days_prior > 27 & days_prior <= 60 ~ "Day 28-60"))

# Establish order for days_prior_cat
test_combine$days_prior_cat <- factor(test_combine$days_prior_cat,
                  levels = c('Day 01-07', 'Day 08-14', 'Day 15-20', 'Day 21-27', 'Day 28-60'))
```

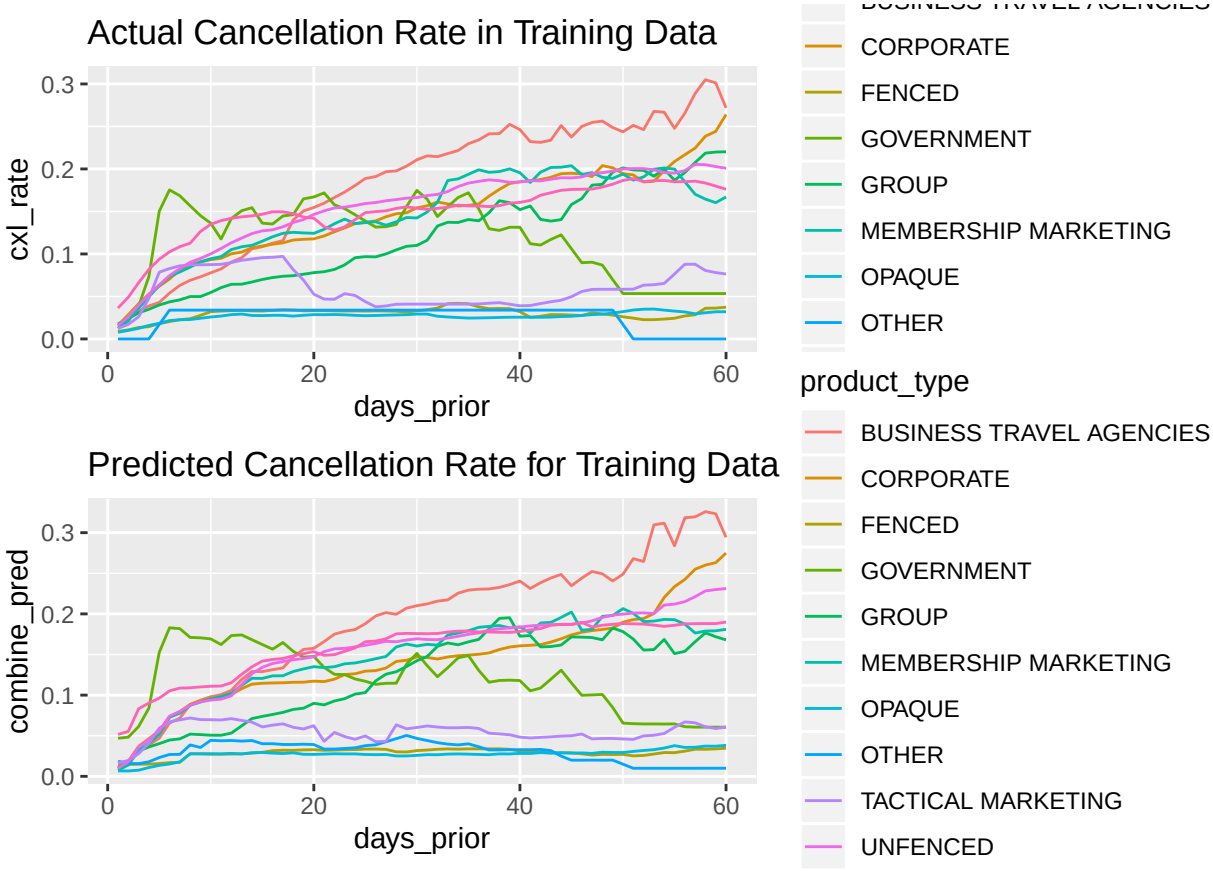
```r
# In-sample
## MAE
MAE_in <- train_combine %>% group_by(days_prior_cat) %>%
  summarise(mae = round(mae(OTB_to_survive,(OTB-OTB*combine_pred)),digits = 4))

## MAPE
MAPE_in <- train_combine %>%
  filter(OTB_to_survive != 0) %>%
  group_by(days_prior_cat) %>%
  summarise(mape = round(mape(OTB_to_survive,(OTB-OTB*combine_pred)),digits = 4))

## MASE
MASE_in <- train_combine %>%
  group_by(days_prior_cat) %>%
  summarise(mase = round(sum(abs(OTB_to_survive - (OTB-OTB*combine_pred)))/
                           sum(abs(OTB_to_survive - naive_survive_pred)),digits = 4))

# Out-sample
## MAE
MAE_out <- test_combine %>% group_by(days_prior_cat) %>%
  summarise(mae = round(mae(OTB_to_survive,(OTB-OTB*combine_pred)),digits = 4))

## MAPE
MAPE_out <- test_combine %>%
```

Table 1: Combine model

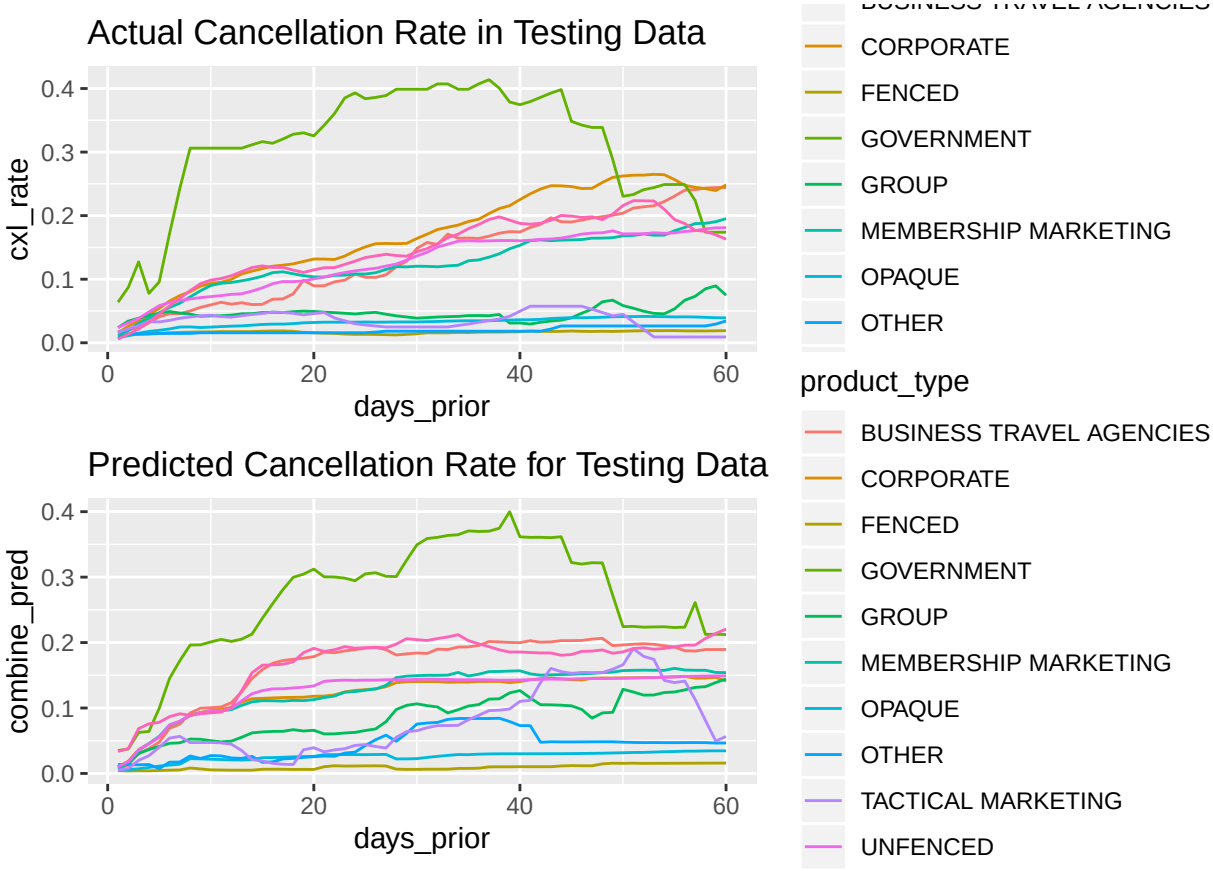
| Days Prior | MAE_in | MAPE_in | MASE_in | MAE_out | MAPE_out | MASE_out |
|---|---|---|---|---|---|---|
| Day 01-07 | 1.4484 | 0.0373 | 0.2120 | 1.8465 | 0.0291 | 0.2103 |
| Day 08-14 | 1.7534 | 0.0559 | 0.4352 | 2.4203 | 0.0467 | 0.3885 |
| Day 15-20 | 1.7383 | 0.0587 | 0.4977 | 2.4820 | 0.0584 | 0.4897 |
| Day 21-27 | 1.5590 | 0.0708 | 0.5160 | 2.3717 | 0.0745 | 0.5266 |
| Day 28-60 | 1.0443 | 0.0990 | 0.4909 | 2.1031 | 0.1246 | 0.5195 |

```
  filter(OTB_to_survive != 0) %>%
  group_by(days_prior_cat) %>%
  summarise(mape = round(mape(OTB_to_survive,(OTB-OTB*combine_pred)),digits = 4))

## MASE
MASE_out <- test_combine %>%
  group_by(days_prior_cat) %>%
  summarise(mase = round(sum(abs(OTB_to_survive - (OTB-OTB*combine_pred)))/
                           sum(abs(OTB_to_survive - naive_survive_pred)),digits = 4))

# Make table
result <- MAE_in %>%
  left_join(MAPE_in, c("days_prior_cat" = "days_prior_cat")) %>%
  left_join(MASE_in, c("days_prior_cat" = "days_prior_cat")) %>%
  left_join(MAE_out, c("days_prior_cat" = "days_prior_cat")) %>%
  left_join(MAPE_out, c("days_prior_cat" = "days_prior_cat")) %>%
  left_join(MASE_out, c("days_prior_cat" = "days_prior_cat"))
names(result) <- c('Days Prior', 'MAE_in', 'MAPE_in', 'MASE_in', 'MAE_out', 'MAPE_out', 'MASE_out')

result <- kable(result,
  caption = "Combine model"
)%>%
  kable_styling(bootstrap_options = 'condensed', full_width = F)

result
```

```
library(gridExtra)
grid.arrange(
  train_combine %>% ggplot(aes(x = days_prior, y = cxl_rate, color = product_type)) +
  stat_summary(geom = 'line', fun.y = 'mean') +
  ggtitle('Actual Cancellation Rate in Training Data'),

  train_combine %>% ggplot(aes(x = days_prior, y = combine_pred, color = product_type)) +
  stat_summary(geom = 'line', fun.y = 'mean') +
  ggtitle('Predicted Cancellation Rate for Training Data'), nrow = 2)
```




```
grid.arrange(
test_combine %>% ggplot(aes(x = days_prior, y = cxl_rate, color = product_type)) +
  stat_summary(geom = 'line', fun.y = 'mean')+
  ggtitle('Actual Cancellation Rate in Testing Data'),
test_combine %>% ggplot(aes(x = days_prior, y = combine_pred, color = product_type)) +
  stat_summary(geom = 'line', fun.y = 'mean')+
  ggtitle('Predicted Cancellation Rate for Testing Data'), nrow = 2)
```


Actual Cancellation Rate in Testing Data
cxl_rate
0.4
0.3
0.2
0.1
0.0
0
20
40
60
days_prior
Predicted Cancellation Rate for Testing Data
combine_pred
0.4
0.3
0.2
0.1
0.0
0
20
40
60
days_prior
CORPORATE
FENCED
GOVERNMENT
GROUP
MEMBERSHIP MARKETING
OPAQUE
OTHER
product_type
BUSINESS TRAVEL AGENCIES
CORPORATE
FENCED
GOVERNMENT
GROUP
MEMBERSHIP MARKETING
OPAQUE
OTHER
TACTICAL MARKETING
UNFENCED