Exploration of results

Questions - Do the addition of spatial lags improve the probability or sales model overall? - Probability model: No - Sales Model: Yes -

```
library(tidyverse)
library(h2o)
```

PROBABILITY model data

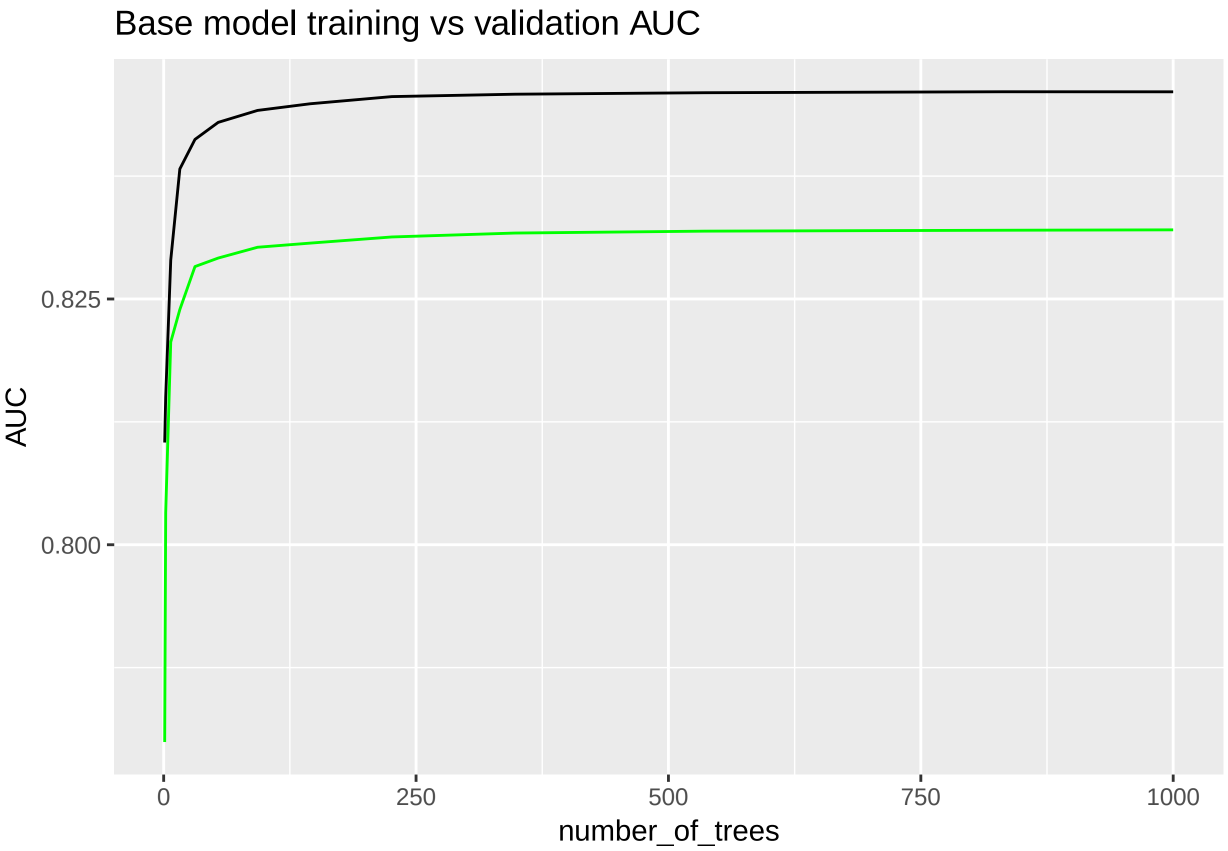
```
prob_base_data <- read_rds("results/prob/p09_prob_of_sale_model_base.rds")
prob_zip_data <- read_rds("results/prob//p10_prob_of_sale_model_zipcode.rds")
prob_radii_data <- read_rds("results/prob/p11_prob_of_sale_model_radii.rds")
prob_evals <- read_rds("results/p15_prob_model_evaluations.rds")

ls()[grep("prob", ls())]
```

```
## [1] "prob_base_data"  "prob_evals"      "prob_radii_data" "prob_zip_data"
```

```
prob_val_metrics <-
  data_frame(type = "Validation"
             , base_AUC = prob_base_data$model@model$validation_metrics@metrics$AUC
             , zip_AUC = prob_zip_data$model@model$validation_metrics@metrics$AUC
             , radii_AUC = prob_radii_data$model@model$validation_metrics@metrics$AUC)

prob_test_metrics <-
    data_frame(type = "Test"
               , base_AUC = prob_evals$base
               , zip_AUC = prob_evals$Zip
               , radii_AUC = prob_evals$Radii)

bind_rows(prob_val_metrics, prob_test_metrics)
```

```
## # A tibble: 2 x 4
##   type       base_AUC zip_AUC radii_AUC
##   <chr>         <dbl>   <dbl>     <dbl>
## 1 Validation    0.832   0.829     0.829
## 2 Test          0.787   0.788     0.796
```

```
prob_base_data$model@model$scoring_history %>%
  filter(number_of_trees>0) %>%
  ggplot()+
  aes(x = number_of_trees)+
  geom_line(aes(y = training_auc), color = "black")+
  geom_line(aes(y = validation_auc), color = "green")+
  labs(title = "Base model training vs validation AUC"
       , y = "AUC")
```



SALES Model Data

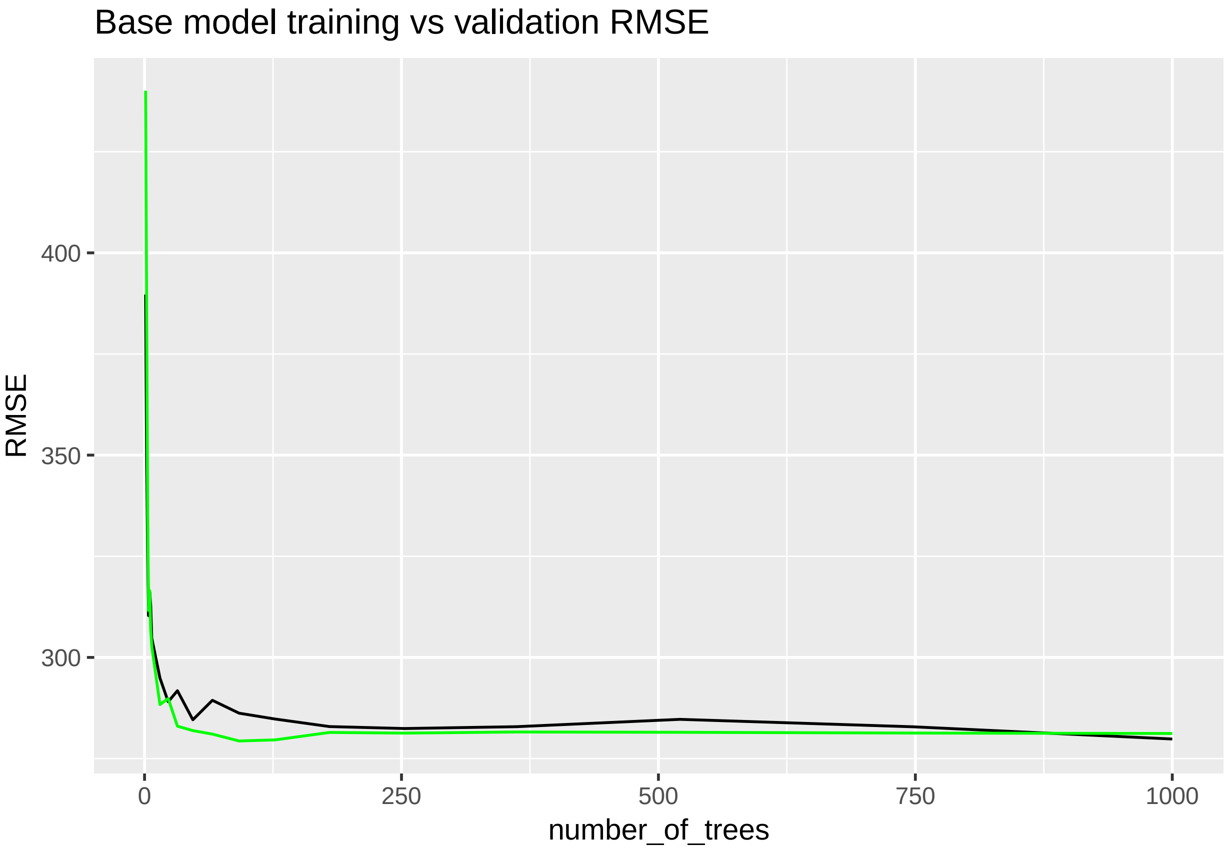
```
sales_base_data = read_rds("results/sales/p12_sale_price_model_base.rds")
sales_zip_data = read_rds("results/sales/p13_sale_price_model_zipcode.rds")
sales_radii_data = read_rds("results/sales/p14_sale_price_model_radii.rds")
sales_evals <- read_rds("results/p16_sales_model_evaluations.rds")
ls()[grep("sales", ls())]
```

```
## [1] "sales_base_data"  "sales_evals"      "sales_radii_data"
## [4] "sales_zip_data"
```

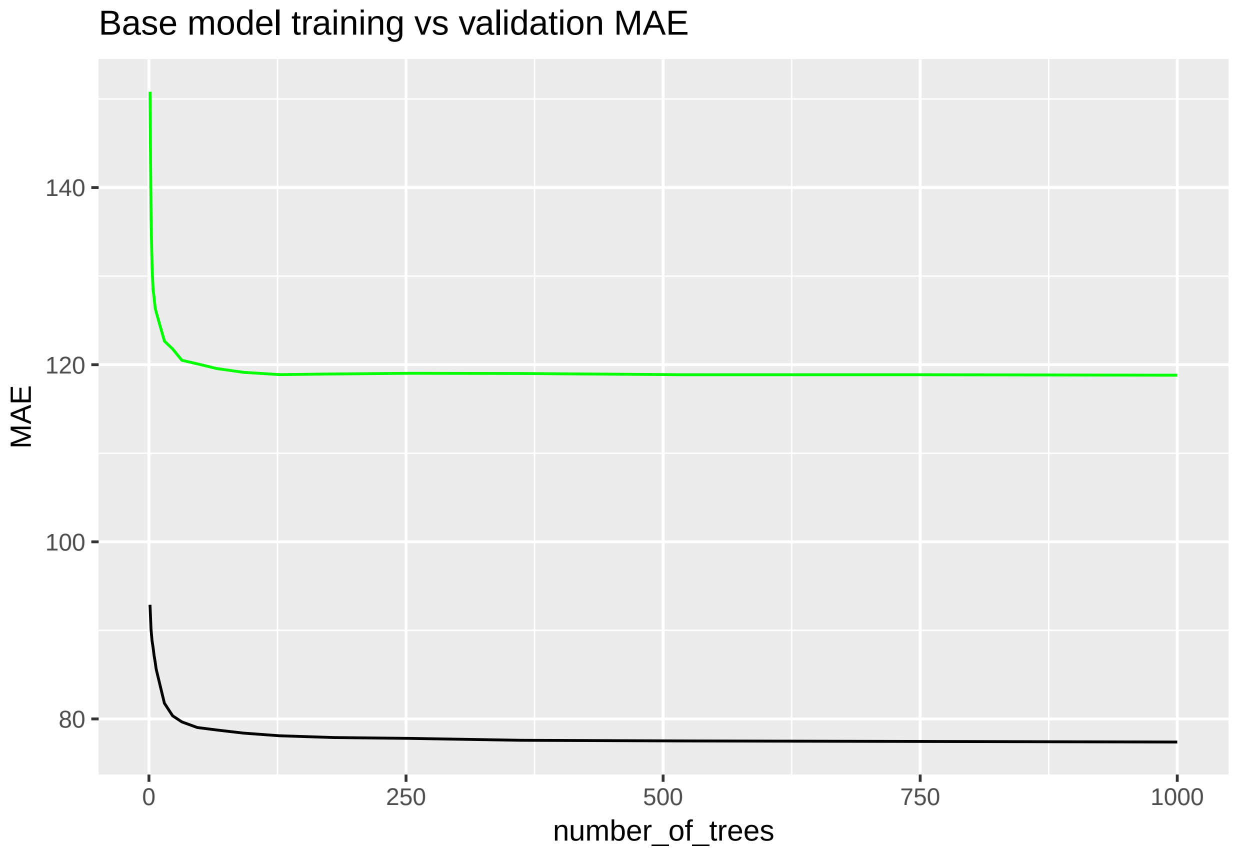
```
sales_val_metrics <-
  data_frame(type = "Validation"
           , base = sales_base_data$model@model$validation_metrics@metrics$RMSE
           , zip = sales_zip_data$model@model$validation_metrics@metrics$RMSE
           , radii = sales_radii_data$model@model$validation_metrics@metrics$RMSE)

sales_test_metrics <-
    data_frame(type = "Test"
             , base = as.numeric(sales_evals[1,"Test_RMSE"])
             , zip = as.numeric(sales_evals[2,"Test_RMSE"])
             , radii = as.numeric(sales_evals[3,"Test_RMSE"]))

bind_rows(sales_val_metrics, sales_test_metrics)
```

```
## # A tibble: 2 x 4
##   type         base   zip radii
##   <chr>       <dbl> <dbl> <dbl>
## 1 Validation   281   302   289
## 2 Test         862  1120   815
```

```
sales_base_data$model@model$scoring_history %>%
  filter(number_of_trees>0) %>%
  ggplot()+
  aes(x = number_of_trees)+
  geom_line(aes(y = training_rmse), color = "black")+
  geom_line(aes(y = validation_rmse), color = "green")+
  labs(title = "Base model training vs validation RMSE"
       , y = "RMSE")
```



```
sales_base_data$model@model$scoring_history %>%
  filter(number_of_trees>0) %>%
  ggplot()+
  aes(x = number_of_trees)+
  geom_line(aes(y = training_mae), color = "black")+
  geom_line(aes(y = validation_mae), color = "green")+
  labs(title = "Base model training vs validation MAE"
       , y = "MAE")
```


Base model training vs validation MAE
MAE
140
120
100
80
0
250
500
750
1000
number_of_trees