Predictive Analytics Homework 2: Prediction in Linear Models

Introduction

This homework assignment will give you practice interpreting this week's concepts. A few exercises require modifying code that is provided.

Completing this assignment will help you be able to:

- Identify the basic components of a *predictive modeling* (aka supervised learning) task.
- Distinguish between *features* and *targets* for a given task.
- Compare and contrast regression tasks and classification tasks, and give examples of each
- Select an appropriate error metric for a supervised learning task (MAE, MAPE, accuracy, etc.).
- Write accurate descriptions of model accuracy in plain language.

Instructions

Create your own Quarto solution file, like last week. This week, your setup chunk should look like:

~~~
```{r}
library(tidyverse)
library(tidymodels)
```
~~~

Suppose we want to predict the sale price of homes (generalizing across different homes, not over time; for better or worse let's assume the market isn't changing very quickly). We have data like what we've been looking at in lecture: characteristics of homes like their square footage and location, and what price they sold for.

Exercise 1: Setting up the problem (6pt)

This exercise is about planning our predictive modeling task; we're not actually looking at the data yet.

a. Name one or more *features* for this task.
b. Name one or more *targets* for this task.
c. Is this a *regression* or *classification* task?
d. What would be an example of an appropriate error metric for this task?
e. Write a sentence that you would use to summarize the expected performance of this model to a decision-maker. (i.e., summarizing how *well* the model works, not *how* it works). Make up reasonable values because you have not yet fit this model.
f. What would be an example of an *inappropriate* error metric? Why would it be inappropriate?

Exercise 2: Fitting and evaluating a model (4pt)

Work with the Ames housing dataset, like we used in the slides. Use this code to load the data:

~~~
```{r}
ames_home_sales <- read_builtin("ames", package = "modeldata") %>%
  mutate(Sale_Price = Sale_Price / 1000) %>%
  filter(Gr_Liv_Area < 4000, Sale_Condition == "Normal")
```
~~~

a. Hold out 10% of homes to validate the model.
b. Fit a *linear regression* model to predict `SalePrice` from `Gr_Liv_Area` on the training set. (*Note: one of the slides examples fits a decision tree; don't get confused.*)
c. Evaluate its MAE and MAPE on the validation set.
d. Write a sentence that you would use to summarize the expected performance of this model to a decision-maker.

You'll find all the code needed for this on the slides. Refer to the Community Resources document for suggestions on how to get help with this.

Exercise 3: Above or Below Median (5pt)

Let's change the problem: let's try to predict, instead, whether a home will sell for *above or below the median price* of all homes in the dataset. That is, for each home, we are tasked to say either "above median" or "below median" (rather than a specific price). **Repeat exercise 1 for this new task.**

Exercise 4: Fit and evaluate the above-or-below-median model (3pt)

Repeat exercise 2 for this new task, using an appropriate type of model and metrics that we've studied this week.

You can use this code to construct the output variable:

~~~
```{r}
ames_vs_median <- ames_home_sales %>%
  mutate(sale_category = case_when(
    Sale_Price > median(Sale_Price) ~ "Above Median",
                               TRUE ~ "Below Median"
  ) %>%
    as_factor() %>%
    fct_relevel("Above Median") # Make sure that "Above Median" is considered the positive cl
  )
```
~~~

> Note: to compute classifier metrics, we need to tell the `yardstick` package which outcome should be considered the "positive". Its convention is that whatever factor level comes *first* is considered positive. The `fct_relevel` changes the levels so that whatever factor is specified comes out first. Let's check that it is indeed the first category now:

~~~
```{r}
levels(ames_vs_median$sale_category)
```
~~~

```
[1] "Above Median" "Below Median"
```

Make sure that you include an evaluation of this model using one or more appropriate metrics.

Notes

One of your evaluations should refer to the concept of a "false positive" or "false negative". Make sure that you describe *what a false positive or false negative means.*

Submitting

Follow the same instructions as for the previous homework.

❗ Important

Make sure that you Render your file before Exporting the HTML.