BAN 502 Course Project

Project Description

For this project you are provided a dataset of home sales in the city of Ames, Iowa. The response variable in the dataset is “Above_Median”. This variable indicates whether or not a home sold above or below the median home sales price in the dataset. Your ultimate objective will be to develop predictive models to predict this response variable. The project features two phases. Each phase is described below.

Phase 1 Project Description

For Phase 1 you will conduct a thorough exploratory/descriptive analysis of the dataset. Please **DO NOT** build any predictive models (e.g., logistic regression, trees, etc.) in this phase.

Assume that your "audience" for this work are non-technical decision-makers.

Phase 1 Deliverables:

There are two deliverables for Phase 1.

Deliverable 1**:** A PowerPoint presentation summarizing your findings from Phase 1. The presentation should be no more than seven slides (including a title slide). Your findings should indicate which variables may be strong predictors of "Above_Median" as well as any other interesting descriptive findings. You should include a charts/visuals in the presentation. There should NO VISIBLE R CODE in this deliverable. As noted above, you should assume that the target audience for the deliverable is relatively "non-technical." NOTE: If you create any variables (i.e., by combining or modifying existing variables), please note this.

Deliverable 2**:** A knitted Word document of your Phase 1 R work.

Submit the deliverables via Canvas.

Phase 2 Description

In Phase 2 you will build predictive models to predict the variable "Above_Median". Your will develop multiple predictive models to predict this variable You should fully document (in your R Markdown file, not in your PowerPoint deliverable) all model building efforts. You should use a training/testing split and may choose to apply k-fold cross-validation when building your model on the training set. Please employ multiple techniques (logistic regression, classification trees, random forests, etc.).

As in Phase 1, assume that your "audience" for this work are non-technical.

Phase 2 Deliverables:

There are two deliverables for Phase 2:

Deliverable 1: A PowerPoint presentation summarizing your findings from Phase 2. The presentation should be no more than seven slides (including a title slide). Your findings should focus on the practical implications of your findings. If your findings are "weak", you should indicate so. You should include appropriate charts/visuals in the presentation. There should NO VISIBLE R CODE in this deliverable. As noted above, you should assume that the target audience for the deliverable is relatively "non-technical."

Deliverable 2: A knitted Word document of your Phase 2 R work

Submit Deliverables 1 and 2 via Canvas.

Hints/Suggestions/Warnings for Phase 2:

- Provide a simple summary table showing your models' performance on the training and testing sets.
- Be careful using predictor variables that have many factor levels (categories), especially if the categories only have a few observations in them. You may wish to use grouping (Hint: We have a "step" function that can be used in your Tidymodels recipe to reduce the number of levels)
- I would probably NOT use Latitude and Longitude as variables in any of your models (unless you severely modify them in advance).