Stat 897 Fall 2017 Project 2

Penn State

Due November 5, 2017

Classification

This project is to be completed individually. You may submit pdf only (Rmd is not needed and you can use another word processing tool if you like).

We will use the Ames Housing dataset, which has 82 variables and 2930 observations (AmesHousing.txt).

Your goal is to classify sales into two classes, those that sold for greater than or equal to USD200,000 and those that sold below. **Exclude the Order, PID, and of course SalesPrice variables from your predictors.** You may want to combine variables (e.g. summing square feet) and perform various manipulations (e.g. transformations for nonlinearity) we have learned about.

You should try at least logistic regression, GAM, LDA, and KNN techniques, but you are welcome to explore more! Please use the following code to define your test set, and its complement the training set.

```
set.seed(7736)
testindices = sample(2930, round(2930/4)) ## train indices are the rest
```

Write up your results in a professional report, like you would present to a client or internal customer for your analysis. **The report should be no more than 4 single-spaced pages long and submitted in PDF format.**

It should include an appropriate analysis of the performance of the models you consder, and the reasons for your final choice of model(s). Include any other details from your analysis that you feel are worthy of mention.

The report should provide sufficient details that anyone with a reason able statistics background could understand exactly what you've done and what you concluded. They should be confident you have not overfit.

Do not embed R code in the body of your report (if you are using rmarkdown, use {r echo=FALSE} to supress the printing of the r code), but instead attach the code in an appendix. The appendix does not count towards the page limit.

Grading criteria (out of 25)

10 points: fulfilling the project requirements.

7 points: a model which makes sense, is predictive, but plausibly not overfit.

8 points: the quality of your report (including: clarity of writing, organization, and layout; appropriate use of tables and figures; careful proof-reading; adherence to report guidelines