Apartment applicants Project Statement

General information for candidates

This assignment has two components. One is a statement of the business problem to be addressed. The other is a list of tasks to be done. Your report will consist of responses to ten specific tasks followed by an executive summary. The audience for the task responses is the examination grading team. Hence, technical language can be used. Each task will be graded individually, so be sure any work that addresses a given task is done within the writeup for that task. The final item in your report is an executive summary written for an audience not familiar with analytics concepts.

This document and the report template indicate the points assigned to each of the eleven components. The total is 100 points. Each task will be graded on the quality of your thought process and conclusions. The executive summary will be graded on the quality of the presentation. Note that a component of the grading of the first ten tasks will also relate to the quality of the exposition, but these sections need not be written as formal reports.

At a minimum you must submit your completed report template and an Rmd file that supports your work. Graders expect that your Rmd code can be run from beginning to end. The code snippets provided should either be commented out or adapted for execution. Ensure that it is clear where in the code each of the tasks is addressed. Your thought process and conclusion for each task should be completely contained within your Word report. The Rmd code should be clear, contain commentary, and support your work.

You may submit other files as needed to support your work. In addition to Word (.docx) and RStudio (.Rmd) files, you may also submit Excel files (.xlsx or .csv). There is a limit of 10 files and no file can be larger than 25MB.

Business Problem

Your client, a property management firm in NYC, needs you to help them attract more tenants to their apartments. This will help them to increase their profitability by signing more leases.

The data is the number of rental applications along with the physical characteristics of the apartment. Some data cleaning was performed in advance. The data dictionary at the end of this document describes the available variables. The target variable *applicants* counts the number of people who filled out a housing rental application within the given month for each apartment building. Only vacant apartments are included.

There are several apartment layouts such as a studio, convertible, one bedroom, one bedroom plus dens, or two-bedroom or three-bedroom residences. Each layout has a different size and number of bathrooms. In addition, each building, depending on whether it is a single-story, residential house or a multi-story apartment high rise, has a different number of units in the building. The residential houses only have 1 unit whereas the high rises can have up to 5 in the same building.

Your client needs these results to be at the per-unit level. The target variable applicants counts the total number of applicants in each building. The other variables are at the per-unit level. For example, the

sale_price is the same for all units in the same building. All units were also given the same rating of interior and external quality, have the same number of bathrooms, and have the same sizes of living area and lot area.

Your assistant has done some preliminary analysis, which is scattered throughout the Rmd file:

- To save you time, when no changes need to be made, there is a comment "No changes needed."
- In places where you need to fill in your own values, uppercase variables are used such as VARIABLE_NAME as placeholders.

Specific Tasks

1. (5 points) Provide a summary of the target variable

    - Provide summary statistics for *applicants* and *num_units*

    - Choose to use either a histogram or a bar plot to visualize the number of applicants and the number of units. Your assistant was not sure of which type of graph to use. These are labeled as "Graph A: Histogram" and "Graph B: Bar plot". Only include one in your report and explain why this type of graph is appropriate.

2. (10 points) Explore the predictor variables *neighborhood_sale_price*, *sale_price*, and *overall_qual*

    Explore these predictor variables, apply any transforms or adjustments necessary, and comment on whether or not you expect each variable will be important in predicting the number of applicants. Decide on whether *overall_qual* should be categorical or ordinal (numeric).

    Note: you don't need to transform *total_sq_feet* or *gr_liv_area* at this time.

3. (5 points) Use insights from the Marketing Manager

    The apartment's Marketing manager has given you info which will be useful for your predictive models. This is based on their experience of leasing apartments for the past 10 years in NYC. Using the below facts, along with the Data Dictionary, engineer 3 additional features and explain your reasoning in non-technical language. You don't need to create graphs or calculate statistics. Take *total_sq_feet, gr_liv_area, tot_bathrooms, lot_area, overall_qual, and full_bath* and combine them with *sale_price* to engineer 3 new features. If you made any transformation to *sale_price*, then use this adjusted variable instead of the original variable.

    - When it comes to demand for apartments, a common saying is "Location, location, location." That is to say that the neighborhood makes a big difference in the price and the number of people who apply.

- Space is expensive in the city. Tenants are always looking for good value in the cost-per-square foot.
- Demand is seasonal. Most apartments that are near universities change hands during the fall from the end of July and August because most leases are for 12 months and people are required to find housing at the beginning of the school year.
- The appearance and overall quality of the building and cleanliness make a big difference. People like newer apartments which are more modern-looking, or older apartments which have been well maintained.
- People like having their own private bathroom. Larger apartments with more residents need to have more half-bathrooms and full-sized bathrooms to be in high demand.

Template code is provided for you to create new features. Replace the right hand side X and Y with other feature combinations which you wish to use. Change the name from "NEW_FEATURE" to a descriptive name.

4. (5 points) Inspect the *garage_type* variables and make any adjustments needed

   Something is wrong with the garage_type variables. Each apartment should have either no garage, one that is attached to the building, a basement garage, a built-in garage, or a detached garage. However, several of the apartments have values of 0 for all garage types. This does not match the data dictionary. Remove these apartments from the data.

   You do not need to check the neighborhood columns as your assistant has already verified that these are correct for all properties.

5. (10 points) Select GLM parameters.

   Your assistant has begun to fit a GLM and needs you to choose a response distribution which is appropriate for modeling the number of applicants. Choose a response family from either binomial, gaussian, Gamma, inverse.gaussian, or poisson, and then complete the following tasks:

   - Explain, prior to fitting the model, why your choice of response family is the best one for this problem.
   - Choose a link function so that the model will be multiplicative.
   - Choose an offset term (if any) so that the model will predict the number of applicants
   - Choose a weighting term (if any) so that the model will predict the number of applicants

6. (10 points) Fit a GLM to predict the number of applicants

   Construct your GLM as designed in task 5 above. Use all predictor variables. The code provided will split the data into training and test sets and fit the GLM on the training data. When you are done, provide the output for the coefficients, p-values, and AIC.

7. (5 points) Use AIC to select features

   Use stepwise selection to simplify your model. Use the forward direction and AIC as the selection criteria. When you are done, provide the following:

   - The model output for the coefficient estimates, p-values, and AIC
   - An interpretation of the graphs of Residuals vs. Fitted, Normal QQ, and Studentized Residuals

8. (5 points) Fit a LASSO

   Your assistant has fit a LASSO model. Note that in R it is not possible to specify the link function when using the glmnet library and so the identity link is used. Use your formula from the GLM in task 6 for the weights and offset terms.

   Your assistant was confused because some of the variables have coefficients which are exactly equal to zero. Before making any changes to the code, provide a statistical explanation as to why certain variables can have coefficients which are zero in a LASSO. Then fit the model and record what these variables are.

9. (10 points) Create a bagged tree model

   Your assistant has set up code to fit eight decision trees. Each tree uses a 20% sample of the training data.

   - Update the formula of the predictions to use bagging and explain how this helps to reduce overfitting.
   - Fit three different selections for the parameters minbucket, cp, and maxdepth and record the log likelihood. Then choose the parameters which have the best log likelihood.
   - Use the graphs for a few of the individual decision trees to identify the top most important variable. This does not need to be a precise estimate but can be just a few sentences. Explain your reasoning based on the plotted trees. You do not need to include the tree graphs in the Word document.
   - Once you have chosen parameters, compare the log likelihood of tree1 to the bagged trees. Is this better or worse?

10. (5 points) Fit a random forest and measure the variable importance

Your assistant has set up a random forest with 400 trees. This uses the same variables as the LASSO model from earlier, except use num_units is the weighting term. Use the results to answer the following questions

- How does the most important variable compare with the model from task 8? What are some possible reasons for these differences (if any)?

- How do the unimportant variables compare with those which had coefficients of zero in the LASSO? If there is a significant difference, why might this be the case? (Ignore the *num_units* as this is different because it is used as a weighting term.)

11. (10 points) Compare model performance

    Your assistant has provided code to calculate a Poisson log likelihood. Compare the performance and comment on the model which has the best result. Comment on 1-2 ways that each model's performance could be improved if you were to do this analysis again.

12. (20 points) Executive summary

    Your executive summary should reflect the information provided in tasks 1 – 12 as relevant to a Sales manager at the property management firm. Your executive summary should include a problem statement, discussion of data, a non-technical overview of the models which you tested, your recommended solution, and a conclusion.

Data Dictionary

Variable	Description
applicants	The total number of people who apply for a lease at that apartment building, including all apartment units.
sale_price	The sale price of each apartment unit.
num_units	The number of units in the apartment building.
year_sold	Year that the apartment building was sold or remodeled
month_sold	Month that the apartment building was sold or remodeled
overall_qual	Rates the overall material and finish of the building on a scale from 1 to 10 with 10 being the best and 1 being the worst.
exter_qual	Rates the external quality of the building on a scale from 1 to 10 with 10 being the best and 1 being the worst.
full_bath	The number of full-size bathrooms in each unit.
tot_bathrooms	The number of bathrooms of each unit.
central_air	Whether or not each unit has a central air conditioning system (1 = yes, 0 = no).
total_sq_feet	Total square feet.
gr_liv_area	Above ground living area in square feet.
lot_area	Lot size in square feet.

Garage Types. Each property should have either an attatched garage, a basement garage, a built-in garage, a detatched garage, or no garage at all.	
garage_type_attchd	1 = Attatched garage
garage_type_basment	1 = Basement garage
garage_type_builtIn	1 = Build in garage
garage_type_detchd	1 = Detatched garage
garage_type_no_garage	1 = No garage
Neighborhoods: Physical locations within Ames city limits. Each apartment building is in one of these neighborhoods.	
neighborhood_br_dale	1 = Dale
neighborhood_brk_side	1 = Brookside
neighborhood_clear_cr	1 = Clear Circle
neighborhood_collg_cr	1 = College Circle
neighborhood_crawfor	1 = Crawford
neighborhood_edwards	1 = Edwards
neighborhood_gilbert	1 = Gilbert
neighborhoodI_dottrr	1= DOTRR
neighborhood_meadowv	1 = Meadow
neighborhood_mitchel	1 = Mitchel
neighborhood_n_ames	1 = North Ames
neighborhood_n_ridge	1 = North Ridge
neighborhood_n_ridge_hghts	1 = North Ridge Heights
neighborhood_n_w_ames	1 = Nowth West Ames
neighborhood_old_town	1 = Old Town
neighborhood_sawyer	1 = Sawyer
neighborhood_sawyer_w	1 = Sawyer West
neighborhood_somerst	1 = Somer St
neighborhood_stone_br	1 = Stone Bridge
neighborhood_swisu	1 = SWISU
neighborhood_timber	1 = Timber
neighborhood_veenker	1 = Veenker
neighborhood_sale_price	The mean sale price for all units in that neighborhood.