

BUS5PA Predictive Analytics – Semester 2, 2020

Assignment 1: Building and Evaluating Predictive Models

Release Date: 28th July 2020

Due Date: Continuous Progress Assessment 31st July 2020 11.55 pm - 28th August 2020, 11.55pm

Weight: 30%

Format of Submission: A report (electronic form) + electronic submission of the R project (scripts) in LMS Site

Objective:

- a) Revise BUS5PA material on predictive modelling
- b) Demonstrate knowledge of data exploration and selection of variables to apply for the predictive models
- c) Demonstrate knowledge of building different types of predictive models using R
- d) Demonstrate knowledge on comparing and evaluating different predictive models
- e) Relate theoretical knowledge of predictive models and best practices to application scenarios

Note: BUS5PA lecture weeks 1-4 have focused on providing you with a foundation knowledge of data preparation for predictive analytics, predictive modelling techniques and implementing these with R. Therefore, this assignment will evaluate your knowledge on building and evaluating predictive models. The deploying of the predictive models and the interpretation of results will be included in assignment 2.

Business Case

An assessor's office in a midwestern state of the USA is in the process to update their method of evaluating the housing prices and intend to use a data driven method to do that. The initial dataset consisted 113 variables describing 3970 property sales in the area between 2006 and 2010. The management is very keen to trial predictive modelling for this task and they have outsourced the task to you.

Part A (10%) – Problem Formulation [Submission Due – 31st July 2020 11.55 pm]

The objective of this section (Part A) is to introduce students to the 'domain understanding and familiarisation' phase data analysts go through prior to the actual analytics. Since you may have to carry out projects in many domains which you may not be familiar with, it is important to be able to develop this ability.

1. Carryout an exploratory study to identify the background and relevant aspects of properties which influence their value, in USA and property evaluation methods?
2. Identify the data sources that would contain information related to the scenario. What is the possible format of information? Will you face any problem in accessing these data?
3. What variables would be useful to build a predictive model to estimate the property price?

To answer these questions, you are encouraged to search for related resources regarding the business case. Following are some examples to refer to:

- <https://www.opendoor.com/w/blog/factors-that-influence-home-value>
- <https://www.usatoday.com/story/money/personalfinance/2013/04/28/24-7-home-features/2106203/>
- <https://www.mashvisor.com/blog/factors-that-affect-property-value/>

- <https://www.hsh.com/homeowner/average-american-home.html>

You are expected to prepare a brief report with the answers to the above questions (Maximum 2 pages)

Part B (40%) – Data Exploration and Cleaning [Submission Due – 14th August 2020 11.55 pm]

Use the provided dataset to answer this section. You are given access to **80** variables that are directly related to property sales from the above-mentioned dataset. Most of these variables are exactly the type of information that an assessor will use to evaluate a property (e.g. when was it built? How big is the lot? What is the size of the living room? Is the basement developed and completed? Number of bathrooms?). You need to answer the following questions with evidence and justifications.

(Note: The dataset will be provided after the submission of Part A.)

1. Which variables are continuous/numerical? Which are ordinal? Which are nominal? What are the methods for transforming categorical variables? Carry out and demonstrate data transformation where necessary.
2. Calculate following summary statistics: mean, median, max and standard deviation for each of the continuous variables, and count for each categorical variable. Is there any evidence of extreme values? Briefly discuss.
3. Plot histograms for each of the continuous variables and create summary statistics. Based on the histogram and summary statistics answer the following and provide brief explanations:
 - a. Which variables have the largest variability?
 - b. Which variables seems skewed?
 - c. Are there any values that seem extreme?
4. Which, if any, of the variables have missing values?
 - a. What are the methods of handling missing values?
 - b. Apply the 3 methods of missing value and demonstrate the output (summary statistics and transformation plot) for each method in (4-a). (hint: the objective is to identify the impact of using each of the methods you mentioned in the 4-a on the summary statistics output above). Which method of handling missing values is most suitable for this data set? Discuss briefly referring to the data set.
5. Evaluate the correlations between the variables.
 - a. Which variables should be used for dimension reduction and why? Carry out dimensionality reduction.
 - b. Explore the distribution of selected variables (from step 5-a) against the target variable. Explain.

Part C (50%) – Building predictive models [Submission Due – 28th August 11.55 pm]

1. Regression Modelling (20%)

- a. Build a regression model with the selected variables.
- b. Evaluate the regression model and carry out feature selection to build a better regression model. You need to try out at least 3 regression models to identify the optimal model.
- c. Compare these regression models based on evaluation metrics and provide the formula for each regression model.

2. Decision Tree Modelling (20%)

- a. Build a decision tree with the selected variables.
- b. Evaluate the decision tree model and carry out pruning to build a better decision tree model. You need to try out at least 3 decision trees to obtain the optimal tree.
Compare these decision tree models based on evaluation metrics and provide the tree plot for each model and explain the outputs.

3. Model Comparison (15%)

- a. Why do we need to build several models in both regression and decision trees (as requested in question 2 and 3)?
- b. Compare the accuracy of the selected (optimal) regression model and (optimal) decision tree and discuss and justify the most suitable predictive model for the business case.