Project Description

Due date. See syllabus section **Tentative Lecture Outline**

Submission process. Each team will submit the following three items as final project package: (1) A list of `PID` of the 20 houses as recommended by your predictive analytics, in a separate Word document dubbed e.g., "Final Recommendation by <*your group members' name*s go here>", (2) final project deliverable, and (3) your entire project folder in a compressed (i.e., zip) file. Besides, you will submit peer evaluations about your other group members anonymously by responding to an email prompt that your Professor will send to you.

Introduction

This project will give you an opportunity to apply many of the predictive analytics techniques covered in class. In your project, you will have a chance to show how well you can apply the techniques you have learned so as to make good decisions. These decisions will be based on your own judgment, which will be informed by the exploratory analyses you conduct, any problem-fixing endeavors you apply on the dataset, and the models you develop and tune.

Objective

The main goal is for your group to recommend for investment the purchase of twenty (20) most luxurious (i.e. priciest) homes out of 100 candidate properties for which the sales prices are not known. The data you will use for this project are described below. A good purchase recommendation is for a shortlist of the 20 homes as close as possible to a final list based on the target prices of the homes; meanwhile, the rank order of homes in your recommended list will not matter, only thing that matters as far as your project's quality of predictive analytics goes is the *number* of homes in your recommendation list that overlap with a list based on the actual target prices. The data you will use for predicting house prices are described below. Before the end of the course and after the submission of your group projects, your professor will disclose to you the actual shortlist and the target prices of each property in the Score data set.

The Data

For data mining purposes, you are given ***four*** (4) ***raw data files*** containing information from an anonymous United States city assessor's office that is located in the North West region. The values in the data files are for individual residential real estate properties sold in that city over a time period of 4 years (year of sale not included in the scope of this project). Descriptions of the variables and the different data files are given below.

You are also given ***one*** (1) ***Score data set*** that consists of many of the variables in the raw data files. There are a couple of caveats regarding the Score data set: a few variables that are in the raw data files are not available in the Score data set, the most critical of which is the `Sale Price` of the houses, along with a few others; there may be missing variables in your Score data set.

Background and the Goal

Imagine that your group is acting as a set of investors who plan to purchase 20 residential properties in the city where the raw data files were collected. The broker of the properties gave your group a list of 100 properties and their information as your Score data set, consisting of many assessed values for them. However, this "Score" data set does not contain the final sale prices or a couple of other variables. For some properties, some values of a few variables may be missing. You will need to develop models and appropriate procedures to predict and rank the `sale price` and recommend the 20 most luxurious houses. The reason for investing in the top luxury houses is that their potential investment growth is believed to be the greatest.

Project Performance Evaluation

Your group's final project performance is to be determined by two parts: (1) quality of predictive analytics as reflected in correctly predicting the top 20 luxury houses, and (2) quality of final project deliverable. Your individual grade of the final project depends on your group's performance set forth as above, as well as your group's peer evaluations of your unique contribution and work ethics.

Criteria of Final Project Deliverable

Your team will manage the entire process included in a Data Mining Project Lifecycle according to the CRISP-DM industry standard, including the deployment phase, but you may readily use the descriptions in this document for the Phase 1 and Phase 2.

Your deliverable will be a detailed report (Word doc) consisting of the following components:

To begin with in your deliverable, in the Business/Research Understanding Phase (Phase 1), you may simply adapt the sections of introduction, data, background and the goal of this description document.

As for the Data Understanding Phase (Phase 2), you may copy-and-paste the variables' description section found below in your deliverable; besides, your Phase 2 must also include: (1) Your data management procedures in SAS EG for the purposes of importing/ appending/ merging/ joining the several raw data files into *one single SAS data set* (tutorials and other useful resources on how to use SAS EG to accomplish these tasks are provided in the Group Project page on Blackboard) to be used as the *train data source* in SAS EM software; (2) A table containing the names and

the configuration of critical properties of the final variables, including the "role" and "level" of each variable; and (3) Document the procedures and findings of any major exploratory data analyses (EDA, e.g., descriptive analysis, correlation analysis, plots and graphs, etc., performed in SAS EG, and/or StatExplore, GraphExplore, MultiPlot, etc., performed in SAS EM) that are deemed important. Analyze and report any notable findings, and denote any issue about the dataset that should be critical to your project, so that you may consider fixing or using any workaround about these issues in the next phases.

Data Preparation Phase (Phase 3) of your deliverable should document any necessary steps (i.e., SAS nodes) and corresponding property configurations in further preparing your single train data set for the next data mining modeling phase, including but not limited to, data partition, replacement (if any), transformation (if any), impute (if needed).

Modeling Phase (Phase 4) and Evaluation Phase (Phase 5) of your deliverable should include all notable data mining model procedures by using and configuring the relevant SAS model nodes. Describe any notable rationale for why you choose to use any model node, as well as how and why you adopt the specific properties of these model nodes; properly document any major lessons learned in the process of calibrating the models. Also, clearly demonstrate what criterion (or criteria), how, and why you choose the criterion or criteria to choose the best model; include any necessary and informative screenshot (must be legible) to showcase what you have done. Generally, your main goal of Phase 4 is to improve your final model performance as much as possible (i.e., *Average Squared Error* on the *Validation* set), whereas your Phase 5 will settle on the final model by comparing your competing candidate models — i.e., a very short list of models with comparable predictive performances — on their advantages, disadvantages, and model complexity, for the goal of choosing the *best model with the easiest interpretability of results in business-term and the least amount of complexity*.

In the Deployment Phase (Phase 6), you will use the appropriate SAS EM procedures to score the `Sale Price`, rank the houses in descending order of predicted prices, and recommend the 20 houses on the top.

Bonus

The Top-5 properties with the highest sold price in the Score data set became the city's "property of the years". Your group will receive bonus points if your recommended list includes these properties, the more the better!

Tips and Suggestions

Exploit the Score Rankings Charts to compare the performance of competing models; Score Rankings Chart may also be useful for identifying important variables.

A useful approach to improving model performance is to focus on important variables instead of blindly accepting all of the variables when conducting kNN, regression, and neural networks. Those variables with clear and smooth distributions and without too many levels unintuitive to interpret may be the ones that you want to consider focusing on; however, those variables whose distributions exhibit unintuitive turns, shapes, spikes, or too many levels may be the ones that you want to consider excluding from certain models, or modifying by binning certain values or levels together to reduce the number of levels or dimensions. Be intelligent and use your creativity in analyzing and tuning your models in order to improve their predictive power!

In each type of method chosen, find the best performing model. If you have doubts, always refer to course materials posted on Blackboard and consult your professor and/or TA. Note that, running an algorithm implies more than connecting the dataset to the respective node and copy-pasting output. You would have to decide which variables may be more useful and which may not be; many of these decisions are tough. Undoubtedly however, the more challenging judgments you make, the closer you may be to the best performing model!

Clearly identify the metric/criteria you use in comparing and evaluating model performance across different methods. Provide rationale as to why you chose such criteria.

Since you will be running many models with many settings, the complexity of the analysis is likely to explode. At each step, document what you have been doing and provide rationale of why you did that. However, not each graph or table would be relevant to forming a cohesive report. Here, you would have to use your judgment in balancing the tradeoff between detail and parsimony of presentation. At the same time, make sure your report is understandable independently, i.e., your narratives should allow me to construct a clear picture of your project.

The few input variables that the Score data set lacks may be of predictive utility; you may want to consider calculating or predicting the values of these variables first, before you proceed to predict `sale price`.

Detailed descriptions of data files and variables are provided on the following pages.

Variables Description

1. PID (Property Identification) – a unique number to identify each property.

2. Lot Area – The size of the lot, measured in square feet, on which the house is located.

3. Lot Shape – The general shape of the lot. A lot with a regular shape has a value of 1, and another with not a regular shape has a value of 0.

4. Bldg Type (i.e., Building Type) – This describes the type of home in terms of its footprint. A single-family detached type of home is indicated by a value of 1, and a townhouse type of home is indicated by a value of 0.

5. Overall Quality – This is a rating of the overall material and finish of the house. The numeric scale of this rating is as follows.

   10 - Very Excellent
   9 - Excellent
   8 - Very Good
   7 -Good
   6 - Above Average
   5 - Average
   4 - Below Average
   3 - Fair
   2 - Poor
   1 - Very Poor

6. Overall Condition: This is a rating of the overall condition of the house. The numeric scale of this rating is as follows.

   10 - Very Excellent
   9 - Excellent
   8 - Very Good
   7 - Good
   6 - Above Average
   5 - Average
   4 - Below Average
   3 - Fair
   2 - Poor
   1 - Very Poor

7. Exterior Quality – This is a rating of the quality of the material on the exterior. A good quality is indicated by a 1, and an average quality is indicated by a 0.

8. Foundation – This describes the type of foundation upon which the house is built. A concrete foundation is indicated by a value of 2; a cinder-block foundation by a value of 1; and brick foundation by a value of 0.

9. Year Built – This describes the year when the house was constructed.

10. Year Remodel – This describes the year when the house was remodeled. If the house was never remodeled, then the "year remodel" is the same as the "year built."

11. Veneer Area of Exterior Wall – This describes the area in square feet of the exterior wall that is veneer.

12. Bsmt Finish Type (Basement Finished Type) – This indicates whether a home's basement is finished or not in the sense that it can be lived in or not. When it is finished, it has a value of 1, and a value of 0 otherwise.

13. Basement Finished Sqr ft – This is the measure of the area of a finished basement.

14. Basement Unfinished Sqr ft – This is the measure of the area of an unfinished basement.

15. Total Bsmt Sqr ft – This is the measure of the total basement area.

16. Heating QC (Heating Quality Condition) – This is a measure of the rating of how well the heating unit is for a house. The rating scale is as follows.
    3 - Excellent
    2 - Good
    1 - Average
    0 - Fair

17. 1st Flr Sqr ft (First floor Sqt ft) – This is a measure of the living space on the first floor of a house.

18. 2nd Flr Sqr ft (Second floor Sqt ft) – This is a measure of the living space on the second floor of a house.

19. Above Ground Living Area – This is a measure of the living space of the entire house, excluding the basement.

20. Number Full Bath Bsmt - This indicates the number of full bathrooms in the basement of a house. A value of 1 indicates that there is a full bathroom and a value of 0 indicates that there is not a full bathroom in the basement.

21. Half Bath House - This indicates whether there is a half bathroom in the house (excluding the basement). A value of 1 indicates that there is a half bathroom and a value of 0 indicates that there is not a half bathroom in the house.

22. Number Full Bath House - This indicates the number of full bathrooms there are in the house, not including bathroom in the basement.

23. Bedroom Above Ground - This indicates the number of bedrooms there are in the house, not including the basement.

24. Number Room Above Ground - This indicates the number of rooms there are in the house, not including the basement.

25. Fireplaces – This indicates the number of fireplaces there are in the house, not including the basement.

26. Garage Type – Whether there is a garage of a given type is described and indicated as follows.
    3 - Attached to house
    2 - Built-In (Garage part of house - typically has room above garage)
    1 - Detached from home
    0 - No garage

27. Garage Cars – This indicates the number of cars that can be accommodated in the garage of the house.

28. Garage Area – This is the size of garage in square feet.

29. Wood Deck Sqr ft – This is the size of the wood deck area in square feet for a house.

30. Open Porch Sqr ft - This is the size of the open porch area in square feet for a house.

31. Sale Price – This is the sale price of a house **(not included in the Score data set)**.

Data files Description

1. Property Survey – 1 → Contains 600 rows

   Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type

2. Property Survey – 2 → Contains 1770 rows

   Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type

3. Quality Assessment

   Variables: PID (Property Identification), Overall Quality, Overall Condition, Exterior Quality, and Foundation

4. House Feature

   Variables: PID (Property Identification), Year Built, Year Remodel, Veneer Area of Exterior Wall , Bsmt Finish Type, Bsmt Finish Sqr ft, Bsmt Unfinish Sqr ft, Total Bsmt Sqr ft, Heating QC, 1st Flr Sqr ft, 2nd Flr Sqr ft, Above Ground Living Area, Number Full Bath Bsmt, Half Bath House, Number Full Bath House, Number Bedroom Above Ground, Number Room Above Ground, Fireplaces, Garage Type, Garage Cars, Garage Area, Wood Deck Sqr ft, Open Porch Sqr ft, Sale Price

5. Score Data - No Sale Price

   Variables: PID (Property Identification), Lot Area, Lot Shape, and Bldg Type, Overall Quality, Overall Condition, Year Built, Year Remodel, Veneer Area of Exterior Wall , Bsmt Finish Type, Bsmt Finish Sqr ft, Bsmt Unfinish Sqr ft, Heating QC, 1st Flr Sqr ft, 2nd Flr Sqr ft, Above Ground Living Area, Number Full Bath Bsmt, Half Bath House, Number Full Bath House, Number Bedroom Above Ground, Number Room Above Ground, Fireplaces, Garage Type, Garage Cars, Garage Area, Wood Deck Sqr ft, Open Porch Sqr ft

The next page shows a graphical representation of the data base.

Graphical representation of the data files