Independent Class Project

Project Description

Due date: see syllabus section **Tentative Course Outline**

Submission process. Each student should submit on Blackboard a single compressed (i.e., zip) package containing two files: (1) a Jupyter Notebook file; (2) a Scored data set (in .CSV format) containing only two columns—`PID` (Property ID) and `PSP` (Predicted Sales Price) for 100 properties. Build your deliverable contents within your project Jupyter notebook file: your codes will be built in the *Code* cells; your deliverable contents will be built in the *Markdown* cells and should include an introduction, project objective, process summary, comments on any EDAs, rationale for doing things in one way vs. another, results interpretation, final conclusion, recommendations, etc.

Introduction

This project will give you an opportunity to apply all of the data analytical techniques covered in class. It will also allow you to indulge in working with a comprehensive data base collected from a real business context. In your project, you will have a chance to showcase and further sharpen your skills on working with Data Frame, Array, and other core data structures of Python, and to show how well you can apply the skills and techniques on a new business problem. You will make your own judgment, some of which will be tough; your judgment will have to be informed by the data wrangling exercises you decide to conduct on the datasets including, but not limited to, exploratory data analyses, visualizations, and any problem-solving endeavors.

Objective

The main goal is for you to properly assemble, clean, and manage a set of raw data files about residential real estate properties; once you have the integrated and cleansed data set, use programming to build two (or more) types of supervised predictive algorithms (at least including `linear regression` and `k-Nearest-Neighbor`; *you are encouraged but not required to self-acquire knowledge and skills of new models beyond the course's coverage*) in order to estimate the sales price of a list of one-hundred (100) houses in a holdout sample. The data you will use are described below.

The Data

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. Descriptions of the variables and the different data files are given below. These raw data files share one common column—`PID` (Property ID), which is a string of 10 numbers denoting the unique IDs of the real estate property; you may want to use `PID` as a "primary key" as if in a database, and join the several data files into an integrated data set (data-frame object). Keep in

mind, there may be missing values, outliers, and/or erroneous values on some of the other variables of these raw data files; your task is to understand every variable, analyze their relationships as you see necessary and appropriate, identify those errors if any, and apply prescriptive measures to fix or work around them.

You are also given one (1) Score data file that consists of all of the variables in the raw data files. There are a couple of critical caveats regarding the Score data set: there may be missing, outlier, and/or erroneous values in several variables; and, the `SalePrice` column is not available.

Background and the Goal

Imagine that you are an investment consultant helping your very important clients purchase residential properties in the city where the Raw data files were collected. You have hand-picked 100 properties that appear to be attracting most of your clients' attention. The broker of the properties has given you the relevant information in your Score data set, including many assessed values of the houses. However, this Score data set does not contain the final sales prices of the properties, because they are not known when you help your clients, are they? Your job is to predict the final sales prices of these homes as close to reality as possible before they even happen! This is so that your clients will not leave too much money on the table by offering too high a price. Specifically, you will need to use the programming skills and techniques that you learned and will learn in the process of doing this project to train and implement at least two (2) data mining algorithms to predict the `SalePrice`.

Project Performance Evaluation

Your final project performance will be determined by two parts: (1) quality of prediction as reflected in the size of the error (mean absolute percentage error or MAPE) of your predicted sales prices, as compared with the actual sales prices, i.e., your quality of prediction is higher when your predicted values are closer to the real values or when the margin of error is smaller, and (2) quality of final project deliverable, whose evaluation criteria are described below.

Criteria of Final Project Deliverable

Build your deliverable contents in the *Markdown* cells of the main Jupyter notebook file.

Your deliverable contents should properly document the procedures, rationales, and findings of the procedures you employ in supporting the objective of the analyses. Your deliverable should at least have ***three*** main sections: (1) a Data Management section including, but not limited to, the procedures undertaken in order to import/ concatenate/ merge/ combine/ join the data files in preparation for data wrangling and munging; (2) an Exploratory Data Analyses section including, but not limited to, any descriptive summary, preliminary analysis, data manipulation

and transformation, outlier and erroneous value diagnosis, missing value detection and remediation actions (such as replacement), and any visualization that can support the decision of properly implementing, tuning, and diagnosing the modeling algorithm(s); and (3) procedures used to choose, train, and evaluate the algorithms, as well as those used to produce a separate Scored Data Frame containing *only two* columns: a `PID` (Property ID) column, and a `PSP` (Predicted Sales Price) column. Do not forget to export this Scored data frame as a .CSV file and submit with your notebook file.

Other Bonuses

The Top-5 properties with the highest sold price in the Score data set became the city's "property of the years". You will receive bonus points if your predicted top-20 priced houses include these properties, the more the better!

Tips on Improving Project Performance

Train different model types, and "diversify" the specifications of models under each type. By leveraging the diversity, you will have a better chance of identifying the best predictive model.

A useful approach to improving predictive model performance is to focus on important variables instead of blindly including all of the variables in your modeling algorithm. 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; meanwhile, those whose distributions exhibit unintuitive turns, shapes, spikes, or with too many levels may be the ones that you want to consider excluding, 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 model(s) in order to improve its predictive power!

Meanwhile, for *k*NN algorithm, choosing an appropriate *k* value is both important and not straightforward. You may want to experiment with a dozen different choices and iteratively seek to evaluate the advantages/disadvantages of them. Be noted that as *k* becomes too large (e.g., > 100), the identified neighbors tend to be far away and "blend in" with the rest of the train set and thus your *k*NN could become less capable of leveraging unique local patterns; on the other hand, as *k* becomes too small (e.g., < 5), the neighbors could be under-representative of the greater pattern existing in the data and thus your *k*NN could become overfitting.

Variables Description

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

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

3. LotShape – 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. BldgTp (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. OverallQuality – 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. OverallCondition: 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. YearBuilt – This describes the year when the house was constructed.

8. YearRemodel – 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."

9. VeneerExterior (i.e., Veneer Area of Exterior Wall) – This describes the area in square feet of the exterior wall that is veneer.

10. BsmtFinTp (i.e., 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.

11. BsmtFinSqft (i.e., Basement Finished Sqr ft) – This is the measure of the area of a finished basement.

12. BsmtUnfinSqft (i.e., Basement Unfinished Sqr ft) – This is the measure of the area of an unfinished basement.

13. HeatingQC (i.e., 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

14. FstFlrSqft (i.e., First floor Sqt ft) – This is a measure of the living space on the first floor of a house.

15. SecFlrSqft (i.e., Second floor Sqt ft) – This is a measure of the living space on the second floor of a house.

16. AbvGrndLiving (i.e., Above Ground Living Area) – This is a measure of the living space of the entire house, excluding the basement.

17. FullBathBsmt (i.e., Number Full Bath Basement) - 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.

18. HalfBathHouse (i.e., 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.

19. FullBathHouse (i.e., Number Full Bath House) - This indicates the number of full bathrooms there are in the house, not including bathroom in the basement.

20. BdrmAbvGrnd (i.e., Number of Bedrooms Above Ground) - This indicates the number of bedrooms there are in the house, not including the basement.

21. RmAbvGrnd (i.e., Number of Rooms Above Ground) - This indicates the number of rooms there are in the house, not including the basement.

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

23. GarageTp (i.e., 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

24. GarageCars – This indicates the number of cars that can be accommodated in the garage of the house.

25. GarageArea – This is the size of garage in square feet.

26. WdDckSqft (i.e., Wood Deck Sqr ft) – This is the size of the wood deck area in square feet for a house.

27. OpenPrchSqft (i.e., Open Porch Sqr ft) - This is the size of the open porch area in square feet for a house.

28. SalePrice – This is the sales price of a house **(not included in the Score data set)**.

Data files Description

1. Property_Survey_1 → Contains 600 rows

   Variables: PID (i.e., Property Identification), LotArea, LotShape, and BldgTp

2. Property_Survey_2 → Contains 1770 rows

   Variables: PID (i.e., Property Identification), LotArea, LotShape, and BldgTp

3. Quality_Assessment → Contains 2370 rows

   Variables: PID (i.e., Property Identification), OverallQuality, OverallCondition

4. House_Feature → Contains 2370 rows

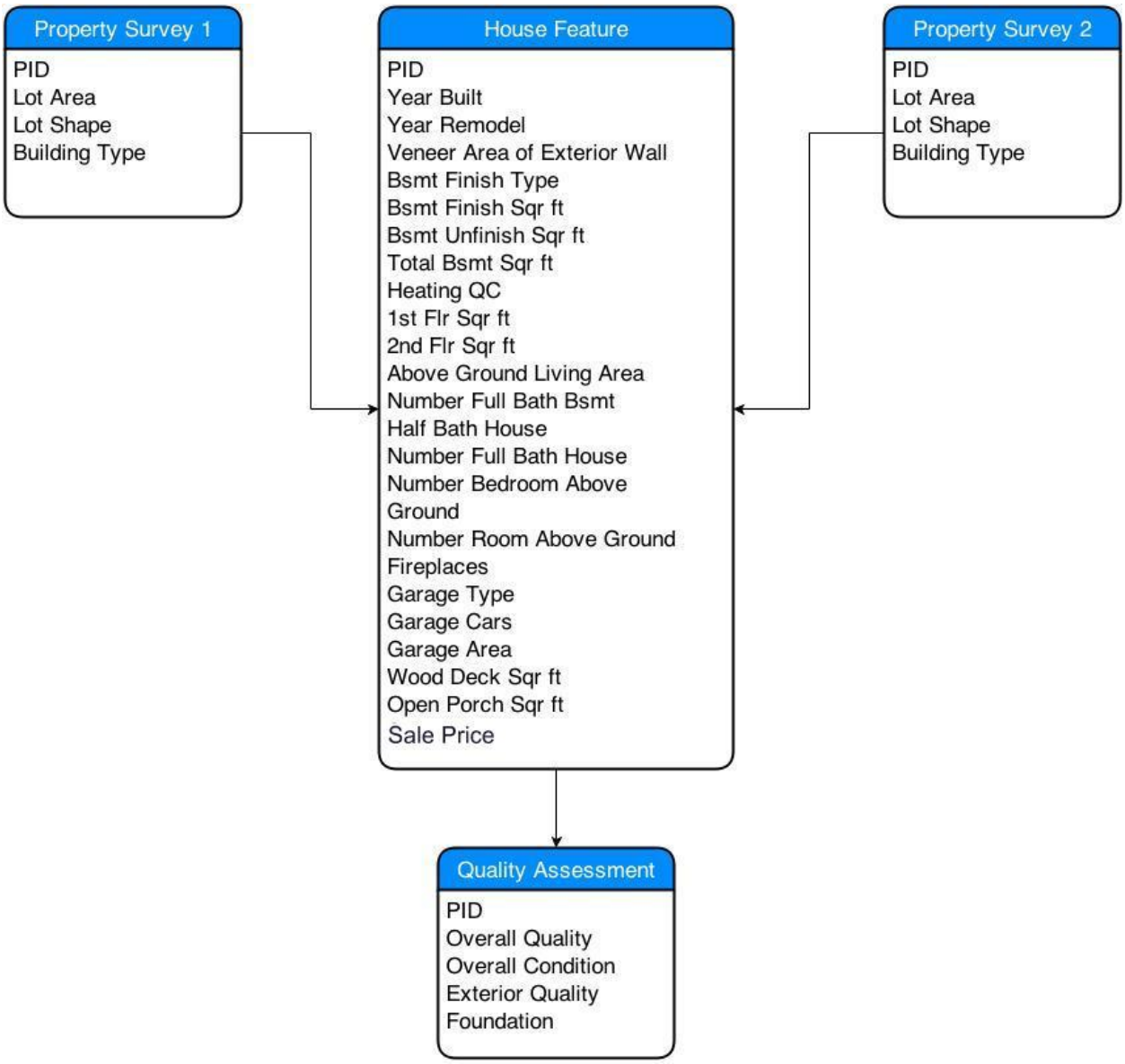
Variables: PID (i.e., Property Identification), YearBuilt, YearRemodel, VeneerExterior, BsmtFinTp, BsmtFinSqft, BsmtUnfinSqft, HeatingQC, FstFlrSqft, SecFlrSqft, AbvGrndLiving, FullBathBsmt, HalfBathHouse, FullBathHouse, BdrmAbvGrnd, RmAbvGrnd, Fireplaces, GarageTp, GarageCars, GarageArea, WdDckSqft, OpenPrchSqft, SalePrice

5. Score_Data – No Sale Price → Contains 100 rows

   Variables: PID (i.e., Property Identification), LotArea, LotShape, and BldgTp, OverallQuality, OverallCondition, YearBuilt, YearRemodel, VeneerExterior, BsmtFinTp, BsmtFinSqft, BsmtUnfinSqft, HeatingQC, FstFlrSqft, SecFlrSqft, AbvGrndLiving, FullBathBsmt, HalfBathHouse, FullBathHouse, BdrmAbvGrnd, RmAbvGrnd, Fireplaces, GarageTp, GarageCars, GarageArea, WdDckSqft, OpenPrchSqft

Graphical representation of the data files

(Shown on next page)


Property Survey 1
PID
Lot Area
Lot Shape
Building Type
House Feature
PID
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
Property Survey 2
PID
Lot Area
Lot Shape
Building Type
Quality Assessment
PID
Overall Quality
Overall Condition
Exterior Quality
Foundation