**Project 1: Predict Real Estate Sale Prices**

**CS6330 Data Science**

**Learning Outcomes**

1. Identify, access, load, and prepare (clean) data sets for a given problem.

2. Determine and apply appropriate experimental setup, evaluation metrics, and models for supervised learning problems.

3. Perform and interpret feature selection to identify relationships between features and predicted variables.

4. Evaluate approaches such as feature engineering, incorporating additional data sets, and model tuning and selection for improving prediction results.

5. Communicate findings through generated data visualizations and reports.

**Overview**

In AI Tools and Paradigms, you explored a real estate transaction data set to learn more about how various characteristics of the properties related to their geographic distributions. In this lab, you are going to focus on developing a regression model to predict the sale price of the properties. You should do your work in one or more Jupyter notebooks but write a lab report summarizing your results.

The dataset contains records for home sales in King County, Washington from May 2014 to May 20215. The data were downloaded from [Kaggle](https://www.kaggle.com/datasets/harlfoxem/housesalesprediction). Here are descriptions of the fields:

* **id -** Unique id for each home sold
* **date -** Date of the home sale
* **price -** Price of each home sold
* **bedrooms -** Number of bedrooms
* **bathrooms -** Number of bathrooms
* **sqft\_living -** Squared footage of the apartments interior living space
* **sqft\_lot -** Squared footage of the land space
* **floors -** Number of floors
* **waterfront -** A dummy variable for whether the apartment was overlooking the waterfront or not
* **view -** An index from 0 to 4 of how good the view of the property was (higher is better)
* condition - an index from 1 to 5 on the condition of the apartment
* **grade -** An index from 1 to 13 where 1-3 falls short of building construction and design, 7 has an average level of construction and design , and 11-13 have a high quality level of construction and design
* **sqft\_above -** the square footage of the interior housing space that is above ground level
* **sqft\_basement -** the square footage of the interior housing space that is below ground level
* **yr\_built -** The year of the house was initially built
* **yr\_renovated -** The year of the house's last renovation
* **zipcode -** What zipcode area the house is in
* **lat -** Latitude
* **long -** Longitude
* **sqft\_living15 - The square footage of interior housing living space for the 15 nearest neighbors**
* **sqft\_lot15 - The square footage of the land lots of the 15 nearest neighbors**

**Instructions**

**Part I: Clean Your Data Set**

In the first part of the project, you are going to through the end-to-end process of training and evaluating a model. This will establish a baseline model you can improve upon in part II.

**Value Representation and Type Unification:** You will first need to clean and prepare the data. All of the entries use value representations, types, and units consistently so there isn't much cleaning to do. You will need to convert some variables (e.g., residential type) into categorical variables. Use the info() and head() methods to explore the data to figure out which columns should be categorical and convert them.

**Clean the Data:** Identify any outliers (e.g., values outside of expected ranges) or missing data. Handle appropriately. Some houses were sold multiple times in the given time range. You’ll see instances of duplicate ids. You can either drop some of the duplicates or average the prices by id.

**Part II: Build and Interpret a Descriptive Model**

1. Build a linear regression model with patsy and statsmodels (excluding the zip, state, and city variables).

2. Assess the fit of the model using the adjusted R2, RMSE, and MAPE metrics, p-value from testing against the intercept-only model, and a plot of the predicted vs true values.

3. Examine residuals to ensure assumptions are met. Perform Shapiro test for normality. Check for patterns described in the lectures.

4. Reduce model by removing unneeded features based on the t-test results. Perform a likelihood ratio test on reduced and full models to ensure that removed features had no overall significant difference.

5. Interpret model coefficients.

**Part III: Build and Evaluate a Predictive Model**

**Experimental Setup:** Initially, you should build a model that predict the total house price given in the "price" column. You will need to set up your experimental conditions. As part of this, you will need to decide an appropriate process for splitting the data into training and testing sets. The goal is for the two sets to have very similar distributions (you want to perform stratification).

Use Numpy's [histogram\_bin\_edges()](https://numpy.org/doc/stable/reference/generated/numpy.histogram_bin_edges.html#numpy.histogram_bin_edges) function with different settings for the bins parameter to bin the samples by price and Seaborn's histplot() method to view the resulting histograms. You should look for a setting that uses fewer bins but keeps the overall shape. Once you are happy with the choice of bins, use Numpy's [digitize()](https://numpy.org/doc/stable/reference/generated/numpy.digitize.html#numpy.digitize) function to assign each record to a bin based on its price. For each bin, randomly assigning 75% and 25% of the records to the two sets, respectively. You can generate a random number between 0 and 1 using random.random(). If the number is < 0.75, put the record in the first set. Otherwise, put the record in the second set.

**Feature Encoding:** For anything you decided to convert to categories, you will need to create dummy variables and remove the original columns.

**Feature Scaling:** Scale the features using the StandardScaler. Make sure the scaler parameters are trained on the training set but applied to both.

**Baseline Model:** In your initial pass, use the LinearRegression or Ridge classes to train a linear regression model to predict the price. Use the model to predict the prices for the test set and evaluate the predictions using the RMSE and MAPE and a plot of the predicted prices against the true prices.

**Other Model Types:** Try other model types such as SGDRegressor with other loss functions, Random Forest Regression, Support Vector Regression (with or without non-linear kernel functions), or multi-layer perceptron (MLP) regression. Note that you will need to tune the hyper-parameters (e.g., C and epsilon for SV Regression) using cross-fold validation.

**Feature Engineering:** You can try engineering other representations of the features that may give better signals. For example, you could try:

* Extract the streets from the addresses and use them as a categorical variable
* Create categorical variables from continuous variables. These categorical versions can be used alongside the original continuous versions – they can be additional variables.
  + Create categorical and/or ordinal forms of the beds and baths variables
  + Bin the sq\_\_ft, latitude, and longitude into discrete bins using the [KBinsDiscretizer](https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing-discretization) transformer.
* 2D binning or clustering of properties by latitude and longitude. You can try defining a 2D grid and use each cell as a bin or using [K-means](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) clustering to cluster the properties. Try different numbers of clusters. Each property should be assigned to one bin or cluster and the associated label should be a categorical variable. Inferring bins and clusters should be done on the training set only but used for assignment of all properties.
* Generating combinations of categorical features such as residential property x number of baths to capture non-linear interactions.

*Hint:* Calculate the residuals and use them to determine which properties have the worst predictions. Explore those records. Do you notice any patterns? Are there features you can engineer that are particularly useful for these samples?

**Add More Data:** Most gains from model performance come from including additional variables from complementary data sets. You will need to find one or more complementary data sets that you can join with the original to add more variables. One example of a complementary data set would be the American Community Survey from the U.S. Census. The ACS gives median household income, unemployment rates, age distributions, and more by zip code. Other data sets might include population densities, crime rates, or school ratings.Once you acquire, clean, and select the desired columns from the chosen data set, you can join it with your current data set.

**Part IV: Report**

Prepare a report describing your results. Include scatter plots of the predicted versus true values and metrics such as RMSEs and p-values to support your points. Particular questions you should answer include:

1. Interpret the various metrics to evaluate the fit and validity of the assumptions for your descriptive linear regression model. How much of the variation is explained by the model? With linear regression, it is assumed that the errors are independent of the magnitudes of the predicted values. Was this true of your model? Can you rely on this model for interpreting the relationships?
2. Which features were significant in the descriptive model? Make a table listing those features, their coefficients, and their units. Provide 1-2 sentences of interpretation for each feature.
3. Evaluate the three approaches for improving the models in Part III. Which approaches were most effective? Which approaches were the least effective?
4. What are some of the limitations of the provided data set?
5. Would you trust the model you created to determine how much to offer when purchasing a house?

**Submission Instructions**

Submit your report as a PDF and notebook(s) as HTML or PDF file(s) through Canvas.