https://github.com/amby-code/UCDA_Ambrose

Project Abstract: Inspired by the Boston data set which was part of the data camp lessons , I looked for a similar real word dataset . The closest I found in the wild was the property price register set-up by an Irish government agency. But this data was limited . However it was augmented by daft.ie the top property website in the country. The daft data added other features to the existing data , like number of rooms . However a Clean the data and perform supervised learning so as to predict house sales for property information. The resulting data set turned out to be too limited to run ML models on but I have included it here to show the my work , Importing and analysing data , something that can be difficult to do with the clean Kaggle datasets. (See PART I).

In order to demonstrate ML, I had to go with a different data set this time from Kaggle.

Its the heart_failure_clinical_records_dataset. And I have broken out separate notebook on this (See PART II)

PART I

Introduction: I spent a long time (too long) deciding on what dataset to choose. But all the sets were limited in one way or another or lacked the features I wanted. I eventually found property price registry for ireland . I chose this project because it has many years of data which I hoped would provide plenty of opportunity to demonstrate cleaning and validating of data.

Dataset:

After the 2008 property crash the Irish housing market was in the doldrums for a number of years. In an effort to being clarity and confidence back to the housing market the government of the day established the “Residential Property Price Register’ using Stamp Duty figures from the Revenue Commissioners. The register is available to the public at www.propertypriceregister.ie

daft.ie is Irelands leading online property website. They aggregate property's for sales from all the auctioneers in the Irish market. Daft have used their wealth of knowledge on property sales to take the Governments limited data and fill in other details such as number of bed rooms, bath rooms and the type of property, where available. I contacted daft to see if they had an API but they responded it was for internal use only. Its this dataset I have used.

It was acknowledged up front that errors were present in the data as it was primarily filed electronically by persons doing the conveyancing of the property on behalf of the purchaser.

The sale price given for a number of houses in the dataset does not represent the full market value of the property sold. These property's are indicates as True or False under the 'not_full_market_price' colum. One reason for this is when a property is part residential and part non-residential (e.g. living quarters over a shop) the Register will only contain information about the residential part. I did a sanity check of some daft address and sales on the official government data sets and they tracked giving me confidence in the daft data. This data set is too big for the UCSSD upload and can be found on the GitHub link.

Implementation Process:
See Jupiter notebook UCDA_Ambrose_Part1_RPPR.ipynb
With no API to access the website , I turned to beautiful soup to scrape the site but found the complexity of the daft site too difficult to complete in the time I had to complete the project.Instead I used a package based on beautiful soup called daft-property-price-register which is an **e**asy way of interacting with the daft property price register. (see ref).

Daft did not have information on all properties on the website so I selected for only properties that had this information. And checked for nulls in the resulting dataset.

Data wrangling : The first step in **p**rocessing the dataframe (ps_df_room) is to converted the date column to a date object with *to_datetime* to allow selecting date ranges later.

I wanted to split out the address block into *County* and *Region* columns and did this with some Regex select instructions. But there were a number of double barrelled names I couldn't resolve , as there weren't that many I decided to drop them.

Reviewing the new County column I see there are some errors of non counties and a typo , So I strip out the wrong county names and correct the typo.

I inspect the remaining columns for errors but they appear ok.

My intuition is there should be a relationship between the number of rooms and the price , so make a scatter plot to see. Its obvious there are more problems with the dates in that some houses have up to 70 bedrooms . I select for properties with more than 11 bedrooms. I also went back onto the daft website and having reviewed the data on daft.ie , its clear that properties over 7 bedrooms tended to be a block of apartments or commercial properties, so will leave out as they amounts to just 120 properties.

Im able to run some simple stats on the price ie
max ,min ,mean,STD

The next scatter plot of the number of bedrooms V price is a lot cleaner and looks like there might be a linear relationship. But a review of the price shows that it's not evenly distributed.

Further studying the relationship between bedrooms and bathrooms , It looks like house with more 3 bathrooms more than the number of bedrooms was a bad entry. ie 4 bed room houses with 15 bathrooms. To fix this will drop any houses that the difference between number of bedrooms and bathrooms is greater than 3

before progressing must convert the categorical values to dummy variables . I went with get_dummies for this and the resulting expanded tables is shown . I used a heat map to check the relationships between the variables . It's crowded , so I ran it again focusing on the features that show the most correlation.

the results from the ML models did not give any prediction or insights

Results:
Results from the machine learning were inclusive as the data set turned out to not be up to the job, but there were insights to be gleamed from the data.
The Box plot price_county_box.png showed clearly the highest prices are to be found in Dublin.

Insights:
1. Prices in Dublin are higher which is probably not surprising as its the capital city , but what is surprising was
2. the country second city Cork was 3rd on the price list behind Co Wicklow. It looks like the influence of Dublin is so great its made the next door county second for the highest prices in property.
3. Dublin and Co wicklow have a greater range of higher prices as seen in the size of the box ie the distance between the 25 and 75 percentile , where as the prices in the rest of the country are flat.The total height contains the middle 80% of the data.
4. From the heat map its looks like Irelands property prices can be broken down into Dublin prices and everywhere else.
5. The the greatest influence on price is the number of bedrooms , which is again probably no surprise but is born out on the heat map correlation 0.43

References:
daft-property-price-register developed by RobertLucey
`pip install daft-property-price-register`

Web address for the web interface : https://ww1.daft.ie/price-register/?d_rd=1

PART II

heart_failure_clinical_records_dataset.

Introduction: My second project uses heart failure data analysis. I chose this subject to investigate as this disease runs in my family and its something I would like to know more about.

Dataset:
See Jupiter notebook UCDA_Ambrose_Part2_Heart.ipynb
I went to Kaggle for my second dataset to be sure I had a robust dataset with enough features to run Ml model on.

The heart_failure_clinical_records_dataset variables are …

Age:
Anemia: A disorder which reduces the ability of blood to carry oxygen.
Creatintine phosphokinase: enzyme used to shuttle energy.
Diabetes : Insulin desensitising
Ejection_fraction: amount of blood ejected from the heart.
High_blood_prerssure: High blood pressure.
Platelets: Measure of the oxygen carving cells
Serum Creatinine: byproduct of muscle action.
Serum_sodium: measure of salt levels
Sex: Male/Female
Time: follow up period , days
Implementation Process:

A review the dataset shows that it is clean, there are no null values by looking at the .info() the .head() shows the categorical column are digitised.

I proceed to split the data set , variable set to X and target set y (death). Because I'm using decision trees the inputs don't need to be scaled

I start with a scatter plot matrix for an overview.

looking at the break out of Sex and Smokers in pie charts.

To look at the correlation between the variables with a heat map, there seems to be a high correlation between sex and smoking.

To pull out the features with the most effect on the target variable (death), I run feature correlation matrix. It shows the The features of interest are *Age*,*Ejection_fraction,Serum Creatinine,Serum_sodium* and *time.*

So I make these variables the X data set, the target y (death).

I use a bootstrap aggregrator or Bagging (sample with replacement) As this is a classification model , ie dead or not dead.

Bagging reduces the variance of individual estimators by using sample with replacement . Trees are grown in parallel with the subsamples

To improve the accuracy I used adaBoost. AdaBoost classifier is a fits the classifier on the original dataset and then fits more copies of the classifier to the same dataset but , weights of incorrectly classified instances are adjusted.this is sequential tree growth with weighted samples.

Results:

The features of interest are Age,*Ejection_fraction,Serum Creatinine,Serum_sodium and time.*

The bagging classifier gave an 0.833333 accuracy result.

This was improved to 0.8705 accuracy by using adaboost.

Insights: 5

1. From the pie_male_female charts we see men were 65% of the cases compared to women only 35% . Heart disease as we know is a higher prevalence amongst men
2. From pie_smokers.jpg chart we see Smokers make up 68% of the patients , showing smoking is a high risk factor.
3. From the feature_correlation_matrix we see not surprisingly age and heart function *Ejection_fraction and Serum Creatinine (a marker of muscle over activity) are indicators.*
4. *The week learners of adaBoost resulted in a model with better prediction.*

References:

Kaggle dataset : https://www.kaggle.com/datasets/rithikkotha/heart-failure-clinical-records-dataset