CSE 587 Spring 2023
Project Phase 3

Team
Name: Pavana Lakshmi Venugopal
UBIT: pavanala
Name: Vaidurya Malathesha
UBIT: vaidurya

Title: Analysis of property for taxes in West Roxbury.

Problem Statement: We will study and examine the information to identify the primary causes for the total value of the property to rise, and then determine the amount of property taxes the owner will have to pay every year.

Link to dataset: https://github.com/reisanar/datasets/blob/master/WestRoxbury.csv

Setup Instructions:

- Environment details**:**
  Python 3.x.x
- Steps to run the application locally.
  - Install Python https://www.python.org/downloads/
  - Open terminal of choice, navigate to the downloaded zip location and cd vaidurya_pavanala_phase3\phase3
  - Create virtual environment
    - python -m venv pav_vai
    - source pav_vai/bin/activate
  - Install the python package requirements in the virtual environment created
    - pip install -r requirements.txt
  - command to run: python -m streamlit run .\streamlit_app.py
  - Local website path : http://localhost:8501

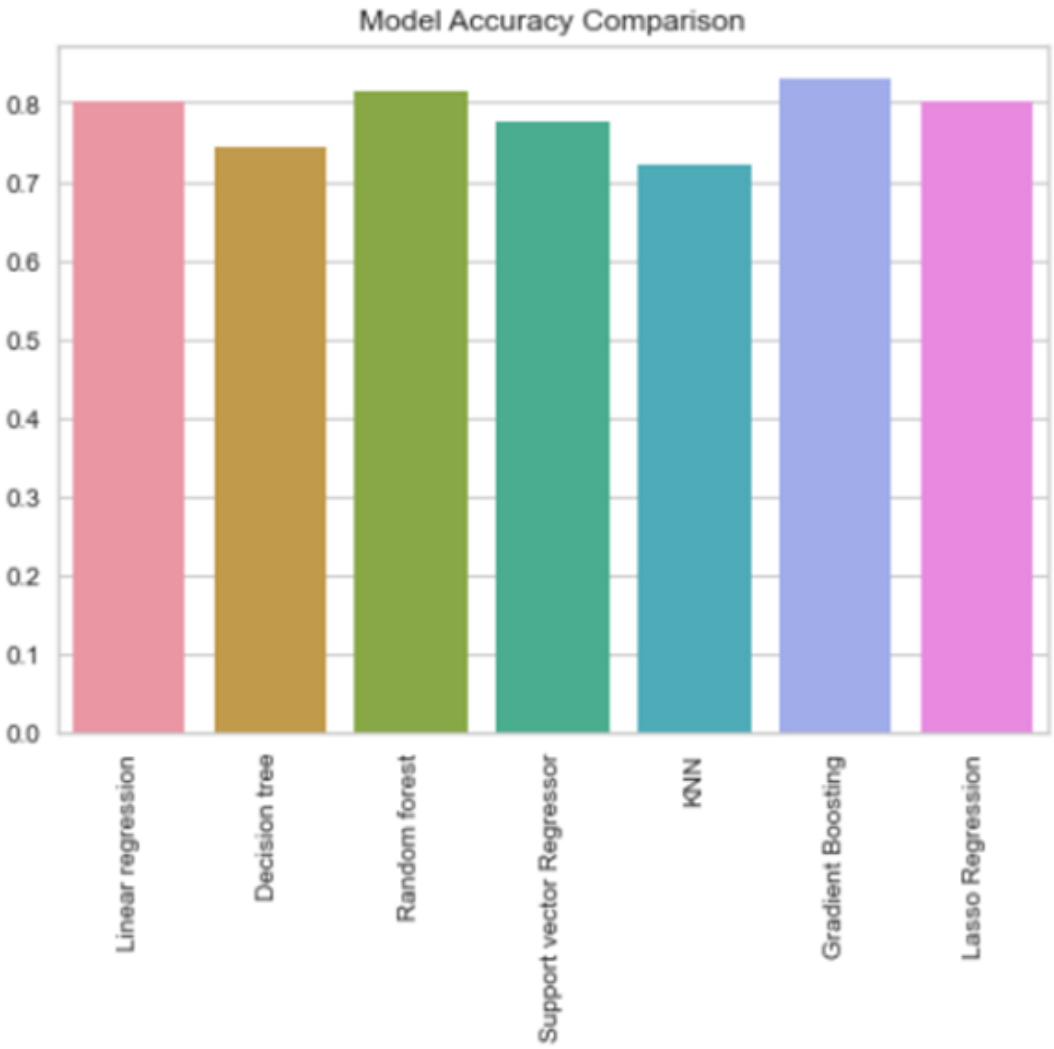
- Steps to Enter new Input data
    - Go to Run your own tab.
    - Enter the data in the available fields.
    - After entering all the values click on “Run Gradient Boosting Model” to get the predicted Total_value.

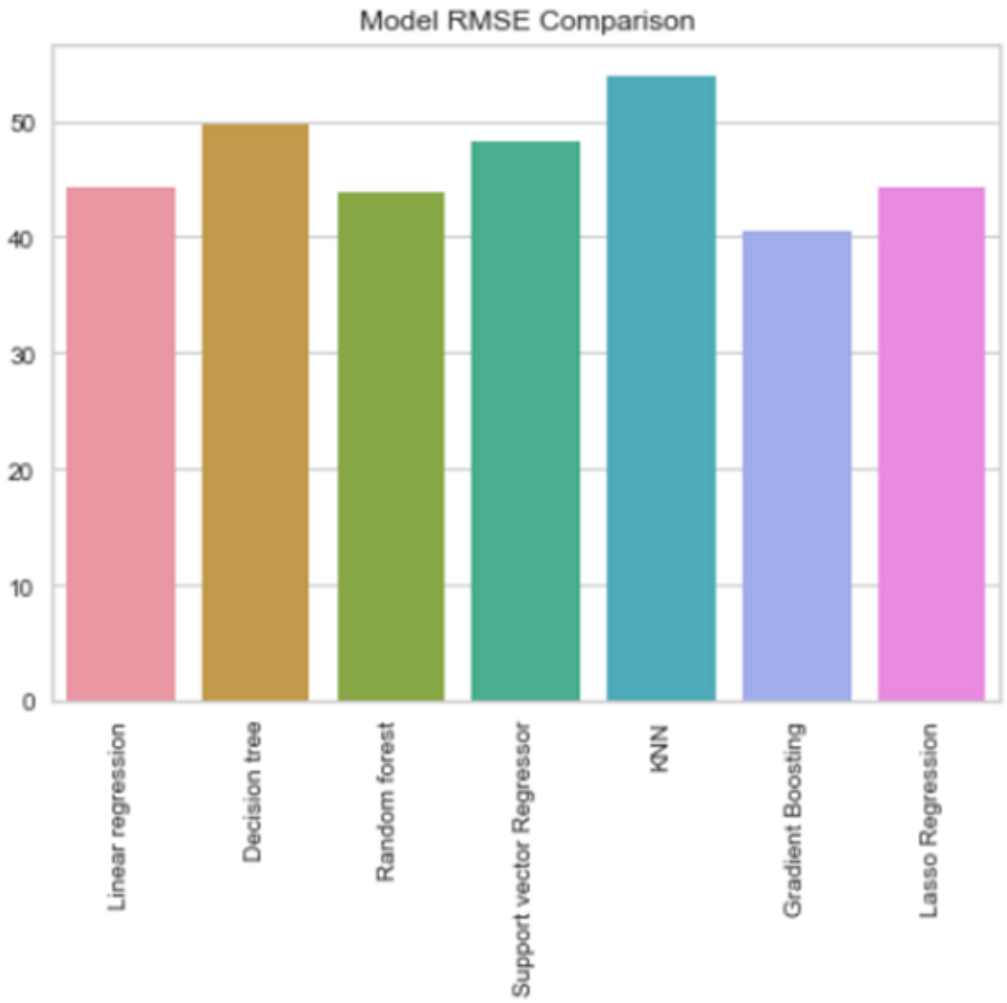
Organization of directories for a project:

```
.
└── vaidurya_pavanala_phase3/
    ├── demo.mp4
    ├── report.pdf
    └── src/
        ├── phase1/
        │   ├── vaidurya_pavanala_phase1.ipynb
        │   └── phase1_report.pdf
        ├── phase2/
        │   ├── vaidurya_pavanala_phase2.ipynb
        │   └── phase2_report.pdf
        └── phase3/
            ├── streamlit_app.py
            └── requirements.txt
```

Model Selection:

In phase 2, we trained five models using our processed data. The results are as shown below.





Though we see close values for Root mean square when compared to Training And testing in all the models, we have seen a huge decrease in error for Gradient descent. Also accuracy being the highest among all. That is 83%. Hence, we choose the Gradient descent model for our prediction task.

Problem Statement Analysis

- A section of Boston, Massachusetts is called West Roxbury. The City of Boston regularly assesses the properties in West Roxbury. When assessing a home, several factors are taken into account such as the year it was constructed, its square footage, and lot size. The final result of an assessment is the 'total value' of the property, which is typically used to calculate the annual property taxes that the owner must pay. Instead of sending assessors out to each property in a city/town/village to calculate assessed values, a more cost effective approach is to have property owners submit the relevant details regarding any upgrades/changes to their homes, and then our models can predict the assessed values based on a combination of historical data and this new data. In other words, it would no

longer be necessary to spend the enormous amount of time required for assessors to visit every home, saving governments hundreds of thousands (if not millions) of dollars.

- The inquiries that will be answered are as follows:
    1. Which data can be retained for Total value?
       Living area is definitely an important factor in determining the total value. This is now evident after Phase1 and Phase 2 operations.
    2. Does any data need modification for our models to interpret?
       Remodel needed to be scaled as machine learning models can only understand numbers to interpret. Hence we flagged it using the get_dummies function.
    3. Which data does not contribute to the predictions?
       In our dataset, GROSS_AREA wouldn't add any significance to the decision making process. And there is LOT_SQFT which would add more weightage than GROSS_AREA in correlating with other features. And the column TAX also doesn't add any value to the predictions. So we have dropped GROSS_AREA and TAX columns.
    4. Which are some of the main reasons for change in total value?
       From our previous phases we know that living area and lot square foot are the main reasons the total value changes.

Recommendations

1. Provides comprehensible visuals - By offering simple to use and understand visualizations, our solution can assist customers in comprehending and interpreting the predictions.This would help users make sense of the data and find patterns that would give them a better understanding of property values.

2. Enable users to make choices that are in their best interests - Our product can empower users to make informed decisions by allowing them to explore the data.They are able to enter various property attributes and view the expected value as a result.This will enable them to make well-informed choices about how to manage existing homes or whether to buy a new home in the West Roxbury neighborhood.

Extending our Project

- It can encompass more regions - Our model is currently focused on WestRoxbury, but we can extend it to other cities and regions.Expanding our model to cover more areas and a wider geographic range will allow more users to benefit from it. It can also help governments in other areas by providing them with data and insights that can be used to improve tax policy.

- Integrate our model into government systems - It would be advantageous to connect our model with government systems so that local governments could utilize it directly to administer property taxes, increasing the usability of our product for users.This would save manpower and resources by reducing the need for assessors to visit every property and check it out.

- Looking for new ways to use our model - Our model is versatile and can be applied to a wide range of problems. For example, in the insurance industry, model like ours can help in underwriting properties or assess risks based on predicted values. We can explore more options, such as these.