|  |  |
| --- | --- |
| Course | Data Analytics Boot Camp |
| Assignment | Project #4: Machine Learning and Data Visualization |
| Team | Group 3: Gurpal, Stephen, Carson, Clarajean, Bilal |
| Due Date | Thursday November 23rd, 2023 |

**Overview**

Prepare a formal 15-minute presentation that covers the following topics:

· Questions that you found interesting and what motivated you to answer them.

· Where and how you found the data that you used to answer these questions.

· Data exploration and cleanup process.

· The analysis, machine learning and visualization process.

· Conclusions: Numerical summary & Visualizations of the summary

· Implications of your findings: what do your findings mean?

**Assignment Requirements:**

* A Python script initializes, trains, and evaluates a model.
* The data is cleaned, normalized, and standardized prior to modeling.
* The model utilizes data retrieved from SQL or Spark.
* The model demonstrates meaningful predictive power at least 75% classification accuracy or 0.80 R-squared.
* The model optimization and evaluation process showing iterative changes made to the model and the resulting changes in model performance is documented in either a CSV/Excel table or in the Python script itself.
* Overall model performance is printed or displayed at the end of the script.
* GitHub repository is free of unnecessary files and folders and has an appropriate. gitignore in use.
* The README is customized as a polished presentation of the content of the project.

**Research Questions to Analyze:**

• What are the historical trends or patterns in housing prices in Texas from 2000 to 2023?

• Based on historical data, can we identify leading indicators that reliably predict future changes in housing prices?

• How does the affordability of housing correlate with changes in median incomes?

• Are there periods where housing became notably more or less affordable for the average citizen?

• How stable is the housing market, and are there indicators of potential instability?

**Datasets to Be Used**

* <https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi>
* <https://www.nber.org/research/data/individual-income-tax-statistics-zip-code-data-soi>
* <https://www.kaggle.com/datasets/blitzr/movehub-city-rankings?select=movehubcostofliving.csv>
* <https://simplemaps.com/data/us-zips>
* <https://www.zillow.com/research/data/>
* <https://www.climate.gov/maps-data/dataset/past-weather-zip-code-data-table>

**Project Description/Outline**

Housing prices in Texas persistently rise alongside increasing incomes. Our team has chosen this topic with the goal of predicting the future changes in housing prices over the years. Through the analysis of housing market data spanning from 2000 to 2023, our Housing Price Prediction Tool aims to forecast whether the housing market in the city of Texas, USA will experience an upward or downward trend in the next two to three years.

**Rough Breakdown of Tasks:**

1. Generate project-specific questions. **Everyone**
2. Obtain relevant datasets. **Bilal**
3. Utilize Jupyter notebook for data import and cleaning. **Bilal**
4. Implement SQL to create our database. **Carson**
5. Deploy AWS to grant everyone access to the database. **Carson**
6. Apply machine learning techniques such as clustering, linear regression, random forest, and gradient boosting. **Stephen and Gurpal**
7. Utilize Tableau to create an informative dashboard. **Clarajean**
8. Compose a comprehensive readme file. **Everyone**
9. Develop our presentation. (we can use Canva) **Everyone**

**Data visualization and analysis summary**: