SPRINGBOARD–DSC PROGRAM
CAPSTONE PROJECT 2: Predict the House Prices
PROPOSAL PREPARED BY: Asiya Shakeel
May,2023

(1) What is the business problem?

Predict the House Prices for residential homes in Ames, Iowa
This is a project on Kaggle where participants are tasked with predicting the sale prices of residential homes based on a dataset of 79 features.
The dataset includes both numerical and categorical features.

(2) Who are the intended stakeholders, and why is this problem relevant to them?

The intended stakeholders are:

1. **Real estate agents and brokers who are interested in accurately pricing residential properties. Accurate predictions of house prices can help them make more informed decisions about listing prices and negotiations with potential buyers.**

2. **Homeowners and prospective buyers who are interested in understanding the value of residential properties. Accurate predictions of house prices can help them make more informed decisions about buying or selling properties.**

(3) Where are the datasets available from?

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/discussion/163335
The dataset for this project is available on Kaggle with the above link.
The training dataset has 1,460 observations with 79 features, and the testing dataset has 1,459 observations with the same features but without sale prices.
Kaggle provides a feature description for each feature in the datasets.

(4) What data science approaches do you anticipate you will use to model the business problem as a data science problem? (*)

This is Supervised Regression Problem because we are trying to predict the price of the house (variable) as a function of the chosen features.

A hybrid approach could potentially be used to combine different data science techniques such as clustering and regression approaches, or classification and regression approaches

(5) How do you anticipate that you will evaluate the performance of each of the data science approaches that you envision?

The performance of each of the models built using multiple regression algorithms can be evaluated using a variety of metrics such as root mean squared error (RMSE), mean absolute error (MAE), Mean Absolute Percent Error (MAPE), and R-squared (R2). In addition to these metrics the following can be used: actual-vs-residual plots, and distribution of residuals with the test set.

Cross-validation techniques can also be used to evaluate the robustness of the models and avoid overfitting.

(6)How do you anticipate that the intended clients will use the results of your CP2 to address the original business problem?

The clients, such as real estate agents, brokers, homeowners, and prospective buyers, can use the results of the "Predict the House Prices" project to accurately price residential properties, make investment decisions, and identify valuable features that have the most significant impact on the sale prices of residential properties.

The results can provide valuable insights and predictions to stakeholders in the real estate industry, which can help them make more informed decisions and achieve better outcomes.

More specifically, interpretability approaches will be used to study the impact of variation of different features on the variation of the target, so the most impactful features can be identified–with respect to increasing/decreasing the value of a property, and also to pose counterfactual analyses to potentially guide improving investments to get maximum return with minimum investment.

> **Deliverables: as required, I will submit all Jupyter notebooks that I will develop, along with a written report, and a presentation slide deck.**

NOTE: algorithms that we will target:

- **Linear Regression >> BASELINE MODEL**
- **Random Forest Regressors**
- **KNN Regressor**
- **XGBOOST**
- **LGBM**

For all of which we will use hyper-parameter tuning as needed