**NIP2 Task 3: Machine Learning Project Proposal**

1. **Project Overview**

Berkshire Hathaway is a leading agent in the Lafayette housing market. The company has new customers daily looking to list their properties for sale. The rapidly changing realty market has made it difficult to price properties before a listing is made. I am proposing a project that will take recorded transaction data for properties that have been recently sold and then use that data to provide a pricing point for new listings generated by the company. The project will utilize sales data from recently sold properties in the area. The program that will be developed will use this information to provide a baseline listing price for a new listing that has certain features based on the sales data gathered for properties with similar features.

Literature review:

**House Price Prediction using Machine Learning** (Bhagat et all, 2023)

Looks at various machine learning models when applied to housing price prediction. Concludes that linear regression gives a maximum accuracy of the machine learning models deployed. According to the results, linear regression provides roughly 85% accuracy compared to those other methods. This article provides the basis that linear regression should be used during the development of the proposed project.

# **The Use of Machine Learning in Real Estate Research** (Choy, L.H.T.; Ho, W.K.O., 2023)

This paper looks at the effectiveness of machine learning when applied to housing predictions. It looks at various types of models that can be employed and concludes that machine learning can provide more accurate price predictions than traditional statistical techniques. This article justifies utilizing machine learning for this project.

# Predicting House Prices using Machine Learning (T.A., 2023)

This article summarizes the steps needed to apply machine learning to the housing market to create a viable product. It describes how data should be handled and processed before being applied to a machine learning model. This insight will provide the project with a pathway to obtain the best data set that will be utilized for the project.

This project will utilize Multiple Linear Regression for the machine learning model that will predict housing prices for potential new listings. The model will take features of a potential property and then calculate the listing price based on properties with similar features. The benefit of this model is that it provides the highest accuracy compared to other models according to the literature. Literature has also shown that when machine learning is applied to housing price predictions it provides a more accurate prediction than traditional statistical techniques.

1. Machine Learning Project Design

The scope of this project will be centralized around the sales data associated with recent sales contained in the database for the area specified: Lafayette, Indiana. Data for sales outside of this established area of study will fall out of scope for the current project. The project will work directly with this data as a baseline for the machine learning model to train. After training, the goal is for the model to be able to deduce the price a property should be listed as based on input data. This data input will include two fields, square footage, and number of rooms. Upon completion, the project will provide a prompt for the user to enter the square footage and numbers of rooms for a potential new listing and then provide a price that the property should be listed with.

The CRISP-DM process will be utilized for data gathering while this project is underway. This methodology consists of a cycle that comprises six stages. The first stage, business understanding, has been outlined in this section as well as the project overview. Stage 2, Data understanding, has already been conducted by the proposal team. Based on the business goals, the appropriate information for the data has been selected by the evaluation team. Stage 3, Data preparation, will be conducted by the team to collect the appropriate information desired for the data set from the selected record repository. Multiple linear regression has been selected from the design team for stage 4 of the CRISP-DM cycle. Up next, the team will evaluate the effectiveness of the model created (stage 5). Finally, with approval from the company, the resulting software will be deployed.

The following tasks are needed for project completion: data collection, model programming, model testing, model adjusting, and deployment. Data collection will begin on 9/18/2023 and is believed to require a large amount of time to gather a data set that is robust for the development and testing of the proposed solution. It is believed that this section will take roughly 3 months for a data scientist to collect the desired amount of data and will conclude on 12/18/2023. At 25 dollars per hour for the data scientist working 40 hours per week, this will cost the company 12,000 dollars for the three months of data collection. After data collection, the team will then move into developing the multiple linear regression model. The proposed project timeline for this development is from 12/19/2023 to 1/15/2024. The development will employ free open-source software like PyCharm for the IDE and the Sklearn library for the machine learning model that will be utilized for the project, costing the company nothing. All equipment needed for the project like computers will utilize current company-owned equipment. Development will be conducted by a single programmer during this one month. At 45 dollars per hour, this stage will cost the company $7200 for one month of development. After the development is completed, the model will be tested. This stage will have a proposed two-month time frame from 1/16/2024 to 3/16/2024. During this time, the test data section will be utilized to test the model and for each test, changes will be made to the model to increase the accuracy of the model. Success will be determined by calculating the R^2 coefficient when comparing the predicted results for a test set of data with actual values for that test set. A separate programmer who was not involved during the development of the project will perform this testing to provide thorough and unbiased testing of the developed software. At 45 dollars per hour for this programmer, this two-month testing period will cost the company $14,400. Total costs for project completion amount to $33,600 for project completion.

1. Machine Learning Solution Design

Hypothesis: Given a data set containing critical information about the recent sales of homes in the Lafayette, Indiana area a machine learning program can use the data set to predict the sale value of a newly listed property. To accomplish this goal supervised learning in the form of multiple linear regression will be utilized. Linear regression is simpler and less complex to implement if you know the relationship between the independent and dependent variables has a linear relationship. However, potential outliers can skew result accuracy.

This application will be developed using Python in the PyCharm Community Edition. The performance of the proposed solution will be measured by the percentage of correct predictions made by the model when given input by a user. A test set from the initial data will be separated to be used as a baseline. The library that provides the machine learning algorithm has a reporting mechanism to test the accuracy of the model. This mechanism will be utilized with the saved test data to see if the model's prediction for sale value matches what the property sold for. Furthermore, sklearn is a library that will be utilized for the development of the multiple linear regression model utilized for the project. This library provides a mechanism that will allow the accuracy of the model to be displayed to the user. This will be used to determine the performance of the solution.

1. Description of Data Set

The source for the data that will be used for this project will come from <https://www.zillow.com/lafayette-in/sold/.> This provides information about over 3,800 recently sold homes in the Lafayette, Indiana area. Due to this, documents and records will be the sole data collection method. The primary advantage of using this type of data collection method is that the data already exists. The only effort required by the development team is to determine the desired information, collect it from the source, and format it within the bounds required for the project. This however provides a slight disadvantage in that it will take a decent amount of time to comb through the data and gather what is appropriate for the project.

The aforementioned Zillow website provides information on the listings that were recently sold in Lafayette, Indiana. Each listing provides the same subset of information concerned with each sale. The following data will be gathered for each entry in the data set provided by Zillow: Address, number of bedrooms, number of bathrooms, square footage, and finally sale price. All of these fields are readily available within the source data and persistent with each record. Therefore, there should be no missing data, outliers, dirty data, or other data anomalies. All of the information being gathered is a matter of public record, therefore no data that will be gathered will be sensitive in any way and no additional rules or procedures will be needed to address this concern.

1. Sources

Bhagat, Ayushi, and Gosavi, Mayuri and Shahasane, Aditi and Mishra, Nandini and Nerurkar, Amit, House Price Prediction using Machine Learning (April 9, 2023). Available at SSRN <https://ssrn.com/abstract=4413863> or <http://dx.doi.ort/10.2139/ssrn.4413863>

Choy, L.H.T.; Ho, W.K.O. The Use of Machine Learning in Real Estate Research. *Land* **2023**, *12*, 740. <https://doi.org/10.3390/land12040740>

1. A. (2022, December 28). Predicting House Prices using Machine Learning. Lindedin.com. Retrieved September 19, 2023, from [https://www.linkedin.com/pulse/predicting-house-prices-using-](https://www.linkedin.com/pulse/predicting-house-prices-using-machine-learning-textifyai)

[machine-learning-textifyai](https://www.linkedin.com/pulse/predicting-house-prices-using-machine-learning-textifyai)