Literature Review
CSCI417 – Machine Intelligence

Hatem Hazem Mohamed	**19104895**
Ahmed Hossam Shawki	**19200095**
Seif Eldin Mohamed	**19105145**
Haneen Ahmed Ezzat	**19106010**
Farah Atef	**19100916**
Dalia Fawzy	**19100802**

Table of Contents

Introduction

Machine learning algorithms are employed in both research and several practical applications. Large structured and georeferenced datasets are now readily accessible thanks to digital technology, making it easier for these algorithms to evaluate and spot trends. Further to generate forecasts that support consumers' decision-making. The goal of this study is to determine the most effective machine learning models for predicting home prices. Every year, the cost of homes rises, necessitating the development of a system that can forecast future home prices.

The ability to estimate house prices can be useful to both the developer and the customer in planning the best time to buy a home. Machine learning techniques have recently been used to estimate home prices, making them comparatively fresh methodological technologies. Machine Intelligence techniques will be used in this study to build a house price prediction model to get the best possible predictions.

Property prices always increase consistently and hardly ever decrease in the long or short term; thus the developer must carefully calculate and choose the best approach when deciding the price of a home. The prediction analysis is one of many methods that can be used to establish the value of the home. Researchers wanted to discover how well the constructed model performed on time series data in this investigation. People who intend to purchase a home will likely benefit from predictions of house prices since they will be able to better organize their finances if they are aware of the price range in the future. Additionally, projections of house prices help real estate investors understand the trajectory of local housing costs.

Project phases

1. **Data visualization:**

Data visualization is a technique that employs a variety of static and interactive visuals within a specific context to aid in the comprehension and interpretation of large amounts of data. The data is frequently presented in a narrative format, which visualizes patterns, trends, and correlations that might otherwise go unnoticed. Data visualization is frequently used to monetize data as a product.

We used some visuals and graphs to plot our dataset to illustrate the relations between our features, such as histograms, boxplots, and Pearson Correlation Matrix.

2. **Data preprocessing:**

Real-world data typically contains noise, missing values, and may be in an unusable format that cannot be used directly for machine learning models. Data preprocessing is a necessary task for cleaning the data and preparing it for a machine learning model, which improves the accuracy and efficiency of the machine learning model.

Our data was already clean, we created new features (ex: age, renovation age) by using simple equations on the existing features.

3. **Simple linear regression:**

Simple linear regression is a regression model that utilizes a straight line to estimate the relationship between one independent variable and one dependent variable. Each variable must be quantifiable.

We used:

Simple linear regression:
Results: 0.491

4. Multiple regression:

Multiple regression is a statistical method for analyzing the relationship between one dependent variable and multiple independent variables. Multiple regression analysis aims to predict the value of a single dependent variable based on the known values of the independent variables. Each predictor value is assigned a weight, with the weights representing the predictor's relative contribution to the overall prediction.

We used:

Multiple regression, stage 1: Selected features (6 features)
Results: 0.512

Multiple regression, stage 2: Selected features
Results: 0.648

Multiple regression, stage 3: all features, no preprocessing
Results: 0.695

Multiple regression, stage 4: all features (including derived features)
Results: 0.698

5. **Ridge and Lasso regularization:**

We used these methods to regularize the data, used among the previously mentioned regression, models and came back with different results.

Lasso regularization: coefficients will be shrunk towards a mean of zero, when a dataset is penalized, less important features are eliminated. The reduction of these coefficients based on the provided alpha value results in some form of automatic feature selection, as input variables are eliminated.
Ridge regularization: Like lasso regression, ridge regression introduces a penalty factor to constrain the coefficients. In contrast to lasso regression, ridge regression utilizes the square of the coefficients.

We used:
Using Ridge Regression:

Polynomial Ridge Regression, 1: alpha=50000, degree=2, all features
Results: 0.791

Polynomial Ridge Regression, 2: alpha=1, degree=2, all features
Results: -3168.943

Ridge Regression, 1: alpha=1, all features
Results: 0.698

Ridge Regression, 2: alpha=100, all features
Results:0.691

Ridge Regression, 3: alpha=1000, all features
Results: 0.648

Using Lasso Regression:

Polynomial Lasso Regression, 1: alpha=50000, degree=2, all features

Results: 0.779

Polynomial Lasso Regression, 2: alpha=1, degree=2, all features

Results: 0.778

Lasso Regression, 1: alpha=1, all features

Results: 0.698

Lasso Regression, 2: alpha=100, all features

Results: 0.698

Lasso Regression, 3: alpha=100, all features

Results: 0.691

6. **Polynomial regression:**

It is a type of regression analysis in which the relationship between independent and dependent variables is modeled using a polynomial of nth degree.

We used:

Polynomial regression, 1: degree=2, all features, no preprocessing

Results: 0.813

Polynomial regression, 2: degree=2, selected features, no preprocessing

Results: 0.714

Polynomial regression, 3: degree=3, selected features, no preprocessing

Results: 0.517

Polynomial regression, 4: degree=3, all features, no preprocessing

Results: -0.9

Polynomial regression, 5: degree=2, all features

Results: -11055.779

7. **K-NN regression:**

KNN regression is a non-parametric technique that intuitively approximates the relationship between independent variables and the continuous outcome by averaging the observations within the same neighborhood.

We used:

KNN Regression, 1: k=15, all features
Results: 0.496
KNN Regression, 2: k=25, all features
Results: 0.487
KNN Regression, 3: k=27, all features
Results: 0.486

Conclusion

In our project we compared multiple machine learning models and their accuracy according to our data.
In the evaluation table, polynomial of second degree (all features, no preprocessing) performs the best. However, we are skeptical of its veracity. we favor **polynomial ridge regression** (alpha = 500000, degree = 2, all features), but other models may also be applicable depending on the situation.

References

Kiyakoglu, B. y. (2019, May 23). *Predicting House prices*. Kaggle. Retrieved January 2, 2023, from https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices/notebook