A logo with text on it

Description automatically generated

**Artificial Intelligence &**

**Machine Learning**

Project Proposal

***Housing Price Prediction***

Rafael Aghashirinov

Table of Contents

[1. Motivation 3](#_Toc166173430)

[2. Learning Task 3](#_Toc166173431)

[2.1. Training Experience 3](#_Toc166173432)

[2.2 Learning Task 3](#_Toc166173433)

[2.3 Performance Measure 4](#_Toc166173434)

[3. Related Work 4](#_Toc166173435)

[4. Plan 5](#_Toc166173436)

[5. Risk management 5](#_Toc166173437)

# 1. Motivation

My motivation for this project is to provide a valuable tool for homebuyers, sellers, and real estate investors alike by predicting housing prices accurately.

The goal of this project is to use machine learning techniques to create a robust predictive model that can estimate housing prices based on relevant factors such as rooms, property size, neighborhood amenities, and so on. By automating this process, we can streamline decision-making in the real estate market and empower individuals with more accurate insights into property valuations.

Through the analysis of extensive housing data, I aim to develop a model that predicts property prices. This knowledge can enable more informed decisions regarding property investments and transactions.

The significance of this project lies in its potential to democratize access to reliable property valuation tools. By providing a transparent and data-driven approach to estimating housing prices, we can help individuals make better-informed decisions about buying or selling homes.

# 2. Learning Task

## 2.1. Training Experience

The data I used for this project was taken from the Kaggle dataset “Housing Price Prediction Data”. The dataset consists of a diverse collection of features, including square footage, bedrooms, bathrooms, neighborhood types, and the year of construction. As it was mentioned in Motivation chapter, the provided dataset holds significant potential for addressing the challenge of predicting housing prices. Within this dataset, each property serves as an illustrative case wherein diverse attributes converge to determine a distinct market valuation. Through the application of machine learning techniques, such as training a model on these cases, the system can differentiate complex connections and recurring patterns that affect housing prices. The diverse array of attributes facilitates a comprehensive comprehension of market dynamics, thus enabling the extraction of insightful conclusions.

## 2.2 Learning Task

The learning objective is centered around instructing a machine learning system to forecast housing prices using a collection of input characteristics drawn from the "Housing Price Prediction Data" dataset, obtainable on Kaggle. This dataset contains diverse attributes such as: square footage, bedrooms, bathrooms, neighborhood types, and the year of construction. The aim is to construct a model capable of reliably approximating the market worth of a property utilizing these attributes.

Given that the dataset comprises both numerical and categorical attributes, preliminary steps like addressing missing data, encoding categorical variables, and standardizing numerical features may be imperative to prepare the data for analysis. Subsequently, we will investigate various machine learning methodologies appropriate for regression tasks, including linear regression, decision trees, random forests, gradient boosting, and neural networks. Through experimentation with diverse algorithms and model configurations, our objective is to ascertain the most efficient approach to predicting housing prices based on the provided dataset.

## 2.3 Performance Measure

For a housing price prediction project, the following three performance measures are relatively straightforward and commonly used:

1. Mean Absolute Error (MAE): This metric calculates the average absolute difference between the predicted prices and the actual prices. It provides a measure of how far off the predictions are from the actual prices.

2. Mean Squared Error (MSE): MSE calculates the average squared difference between the predicted prices and the actual prices. It penalizes larger errors more heavily than MAE, making it sensitive to outliers.

3. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and shares its characteristics but is on the same scale as the target variable. It provides an easily interpretable measure of the typical deviation of the predictions from the actual prices.

# **3. Related Work**

Searching for related work, I found that there are very few papers that specifically tackle the problem of House Pricing Prediction:

<https://www.researchgate.net/publication/350430324_House_Price_Prediction_Using_Machine_Learning_Algorithm>

* This paper shows a study on housing price prediction using various machine learning-regression algorithms, such as: Simple Linear Regression, Ridge Regression, Lasso Linear Regression, Support Vector Regression, Decision Tree Regression and Random Forest Regression, and selects the models with the best accuracy score to predict house prices. The paper evaluates the performance of these algorithms using RMSE (Root Mean Square Error) metric.

<https://medium.com/@manilwagle/predicting-house-prices-using-machine-learning-cab0b82cd3f>

* This project presents solution for predicting house prices in California, motivated by the desire to introduce data-driven decision-making in the real estate industry. The project's goal is to surpass the accuracy of individual predictions commonly based on comparable properties. Utilizing California Census data, the project employs various regression models to predict district-level median housing prices. Evaluation metrics such as RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are considered to determine the model's performance. The project outlines a comprehensive approach to address the challenge of predicting housing prices through advanced machine learning techniques, emphasizing the importance of data-driven insights in the real estate sector.

<https://github.com/Viveckh/LilHomie/blob/master/README.md>

* This GitHub repository contains a project named "LilHomie" that focuses on predicting housing prices using machine learning. The project provides code implementations and documentation on how to preprocess the dataset, train machine learning models, and evaluate their performance for housing price prediction tasks.

# 4. Plan

* **Data Collection:** Obtain the housing price prediction dataset from Kaggle.
* **Preprocessing:** Processing the data, to make it suitable for training.
* **Model Selection:** Experiment with various machine learning algorithms suitable for regression tasks.
* **Model Training:** Train a machine learning model to predict house pricing.
* **Evaluation:** Evaluate their performance using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
* **Final Model Selection**: Select the best-performing model based on evaluation metrics and deploy it for making housing price predictions.
* **Result**: A tool for predicting house pricing based on data such as square footage, bedrooms, bathrooms, neighborhood types, and the year of construction.

# 5. **Risk management**

* **If the primary dataset utilized for housing price prediction exhibits deficiencies in quality, such as missing values, inaccuracies, or outdated information, it's recommended to explore alternative datasets available on platforms like Kaggle, government repositories, or real estate platforms.**
* **If the selected model lacks interpretability or transparency, opting for interpretable models such as linear regression or decision trees may be beneficial. Alternatively, leveraging model explainability tools such as LIME can provide insights into model predictions and feature importance.**
* **There is a risk of overfitting the predictive models to the training data, leading to inflated performance metrics and poor generalization to unseen data. To address this risk, techniques such as cross-validation, regularization, and feature selection should be employed during model development to prevent over-reliance on noisy or irrelevant features and promote robust model performance on unseen data.**
* **External factors such as economic fluctuations, regulatory changes, or unforeseen events (e.g., natural disasters) can impact housing prices and undermine the predictive accuracy of the models. Implementing dynamic model monitoring and updating mechanisms that account for changing external conditions can help mitigate the impact of such factors and ensure the models remain relevant and effective over time.**
* **If the chosen machine learning algorithms fail to converge or yield unsatisfactory outcomes, alternative algorithms that better suit the dataset and problem should be considered. This may entail experimenting with simpler models, ensemble methods, or even delving into deep learning approaches if warranted.**