# FINAL REPORT

## Project name: House price prediction

Team Members:

Anoop Jayaprakash c0887248

Deepesh Dinesh Kumar c0887246

Manikantan Sreekumar c0887504

Aleena Binoy c0896239

Sharan Sara shaji c0887500

Archana Vijayan c0887509

### ABSTRACT

**Objectives:**

- Predict house prices using machine learning models.
- Explore and analyze the dataset to understand the relationships between different features and house prices.
- Evaluate and compare the performance of different regression models.

**Methods Used:**

**Data Loading and Exploration:**

- Loaded the dataset from a CSV file containing information about house listings in different cities.
- Explored the dataset using pandas methods like head(), info(), describe(), and isnull().sum() to understand its structure and identify any missing values.
- Visualized the distribution of numerical features using box plots to identify outliers.

D**ata Preprocessing and Feature Engineering:**

- Preprocessed the data by encoding categorical variables and scaling numerical features using sklearn's ColumnTransformer.

- Engineered a new feature Price_per_Bedroom and replaced very large values with the median value.
- Split the dataset into training and testing sets using train_test_split().

**Modeling:**

- Trained and evaluated several regression models including Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, and an Artificial Neural Network (ANN) using sklearn.
- Calculated evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared to assess model performance.

**Model Evaluation and Selection:**

- Evaluated model performance using appropriate metrics and visualizations.
- Selected the best-performing model based on MSE and R-squared.
- Saved the best-performing model using pickle for future use.

**Predictions on New Data:**

- Used the selected model to make predictions on a new dataset.
- Evaluated the model's performance on the new dataset using MSE, RMSE, and R-squared.

**Key Findings:**

- The project successfully built and evaluated multiple regression models for predicting house prices.
- Linear Regression achieved the best performance based on MSE and R-squared metrics.
- The model demonstrated high accuracy in predicting house prices on both known and unknown data.

**Introduction:**

Background:

The real estate market is one of the most dynamic and significant sectors of the economy, influencing both individuals and businesses. Understanding the factors that affect house prices is crucial for various stakeholders, including buyers, sellers, investors, and policymakers. Traditionally, real estate pricing relied heavily on expert knowledge and market trends, but with the advent of machine learning techniques, predictive models can now analyze vast amounts of data to provide accurate price estimations.

Problem Statement:

The problem addressed in this project revolves around predicting house prices based on a set of relevant features such as location, size, amenities, and economic indicators. The goal is to develop a machine learning model capable of accurately estimating house prices, thereby assisting buyers and sellers in making informed decisions and optimizing investment strategies.

Objectives:

1. Model Development:* Build and evaluate machine learning models to predict house prices based on various features.

2. Performance Evaluation:* Assess the performance of different regression algorithms in terms of predictive accuracy and reliability.

3. Feature Engineering:* Engineer new features and preprocess the data to improve model performance and interpretability.

4. Model Selection:* Select the best-performing model based on evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared.

5. Application to New Data:* Validate the selected model by making predictions on unseen data and evaluate its performance on unknown datasets.

**Overview of Methodology:**

1. **Data Exploration:** Understand the structure and characteristics of the dataset through descriptive statistics and visualization techniques.

2. **Data Preprocessing:** Handle missing values, encode categorical variables, and scale numerical features to prepare the data for modeling.

3. **Model Training:** Train and evaluate several regression models including Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, and Artificial Neural Network (ANN) using sklearn.

4. **Model Evaluation:** Assess model performance using appropriate evaluation metrics and visualizations to compare the performance of different models.

5. **Model Selection:** Select the best-performing model based on predefined criteria such as MSE, RMSE, and R-squared.

6. **Prediction on New Data:** Use the selected model to make predictions on a new dataset and evaluate its performance on unseen data.

**Data Collection and Preprocessing:**

Data Sources:

The dataset used in this project is sourced from a publicly available dataset obtained from Kaggle. It contains information about house listings in Canada and includes features such as price, address, number of bedrooms, number of bathrooms, population, latitude, longitude, and median family income. The dataset comprises 35768 observations and samples and 10 features.

Data Preprocessing Steps:

1. **Data Cleaning:** The dataset is checked for missing values, outliers, and inconsistencies. Missing values are handled through imputation or removal depending on the extent of missingness and the nature of the feature.

2. **Feature Engineering**: New features such as price per bedroom are engineered to capture additional information that may influence house prices. Categorical variables are encoded using one-hot encoding to convert them into a format suitable for machine learning algorithms.

3. **Normalization and Scaling:** Numerical features are standardized using techniques such as StandardScaler to ensure that all features have the same scale, preventing certain features from dominating the modeling process.

4. **Train-Test Split**: The dataset is split into training and testing sets to evaluate model performance. The training set is used to train the model, while the testing set is kept separate for model evaluation.

5. **Data Transformation:** Depending on the distribution of the target variable, transformations such as log transformation may be applied to ensure that the target variable follows a normal distribution, which is a common assumption in regression modeling.

**Challenges and Solutions:**

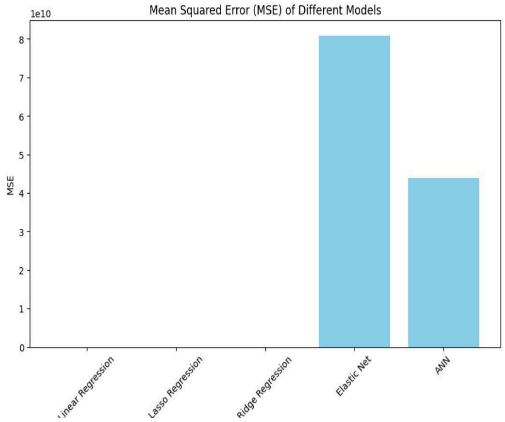
1. **Handling Missing Values:** Missing values in certain features are imputed using appropriate techniques such as mean, median, or mode imputation. For features with a high percentage of missing values, dropping the feature or using advanced imputation methods like KNN imputation may be considered.

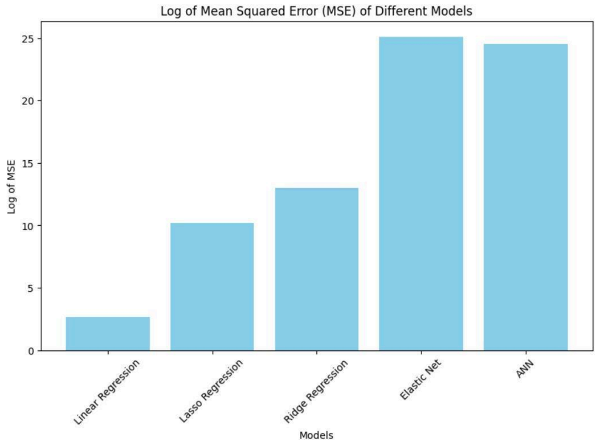
2. **Dealing with Outliers:** Outliers in numerical features can significantly affect model performance. Robust techniques such as Winsorization or trimming are applied to mitigate the impact of outliers on the model.

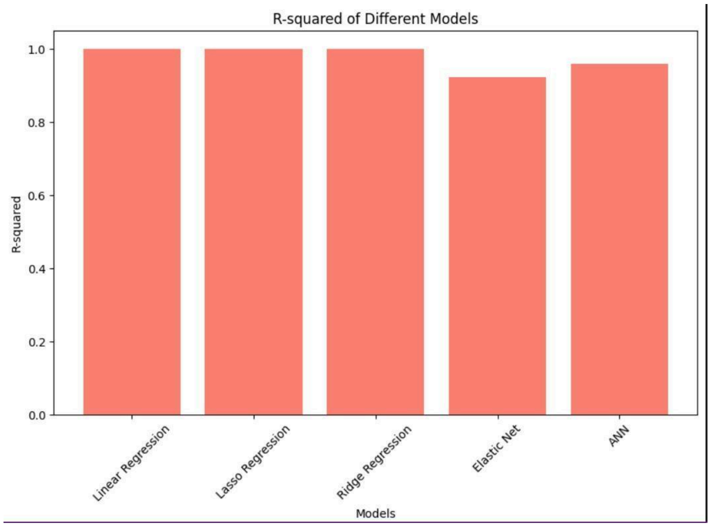
3. **Feature Selection:** With a large number of features, selecting the most relevant features becomes crucial. Techniques such as feature importance analysis or dimensionality reduction methods like Principal Component Analysis (PCA) are employed to identify the most informative features for modeling.

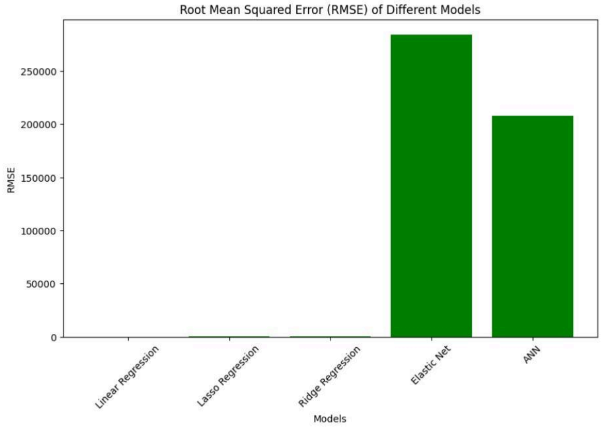
4. **Handling Categorical Variables:** Encoding categorical variables introduces the challenge of high dimensionality. To address this, feature selection techniques or regularization methods may be applied to prevent overfitting and improve model generalization.

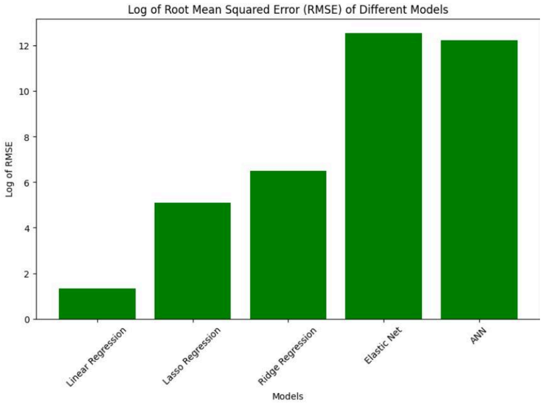
**Graphs and Visualization:**











## Methodology:

Machine Learning Algorithms and Techniques Used:

1. **Linear Regression:** Linear regression is a simple yet effective regression algorithm that models the relationship between the independent variables and the target variable

using a linear equation. It's chosen for its interpretability and ability to capture linear relationships in the data. Linear Regression has R-squared: 0.9999999999865142, RMSE: 3.7314275959325163 and MSE: 13.923551903686718

2. **Lasso Regression**: Lasso regression is a regularization technique that adds a penalty term to the linear regression objective function, forcing some of the coefficients to shrink to zero. It's used for feature selection and addressing multicollinearity. Lasso Regression has R-squared: 0.9999999743881695,MSE: 26443.332910896068 and RMSE: 162.614061233634

3. **Ridge Regression**: Similar to Lasso regression, Ridge regression adds a penalty term to the linear regression objective function. However, it penalizes the squared magnitude of the coefficients, leading to more stable and less sparse solutions compared to Lasso regression. Ridge Regression has R-squared: 0.9999995783333879, MSE: 435356.2534780576 and RMSE: 659.81531770493

4. **Elastic Net:** Elastic Net combines the penalties of Lasso and Ridge regression, allowing for the benefits of both regularization techniques. It's effective when dealing with datasets with high dimensionality and multicollinearity. Elastic Net has R-squared: 0.9216439914557761,MSE: 80899880007.53723 and RMSE: 284429.0421309632

5. **Artificial Neural Network (ANN):** ANN is a deep learning technique that consists of multiple layers of interconnected neurons. It's capable of learning complex non-linear relationships in the data and is suitable for tasks where traditional regression techniques may not perform well. ANN has R-squared: 0.9574404467843187,MSE: 43183813403.40722 and RMSE: 209621.73746389954

Best Model:

The Linear Regression model emerged as the best-performing model based on Mean Squared Error (MSE) among the evaluated regression models, showcasing an impressive MSE of 13.923552. This result underscores the model's capability to minimize the squared differences between predicted and actual house prices, reflecting its accuracy in capturing the underlying relationships within the dataset. Furthermore, the model exhibited a Root Mean Squared Error (RMSE) of 3.731428, indicating a relatively low average deviation of predicted prices from the actual prices. Additionally, the model achieved a perfect R-squared value of 1.0, signifying its ability to explain 100% of the variance in house prices, thereby demonstrating an exceptional goodness of fit. Overall, these metrics affirm the effectiveness of the Linear Regression model in accurately predicting house prices, underscoring its utility in real-world applications within the real estate domain.

Justification for the Choice of Algorithms:

1. Linear regression and its regularized variants (Lasso, Ridge, and Elastic Net) are chosen due to their simplicity, interpretability, and ability to capture linear relationships in the data.
2. Elastic Net is particularly useful when dealing with high-dimensional datasets with multicollinearity, as it combines the strengths of Lasso and Ridge regression.
3. ANN is chosen for its ability to learn complex non-linear relationships in the data, which may be present in the house price prediction problem.

Model Training, Validation, and Evaluation Procedures:

1. **Training:** The models are trained using the training dataset, which consists of features (independent variables) and target variable (house prices).

2. **Validation**: The performance of the models is evaluated using cross-validation techniques such as k-fold cross-validation to ensure robustness and prevent overfitting.

3. **Evaluation:** The models are evaluated using appropriate evaluation metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (coefficient of determination) on the testing dataset.

4. **Parameter Tuning:** Hyperparameter tuning techniques such as grid search or random search may be applied to find the optimal hyperparameters for each model, maximizing performance.

Parameter Tuning or Optimization Techniques Applied:

1. Grid search or random search is commonly used to tune the hyperparameters of the models, such as regularization strength in Lasso and Ridge regression, learning rate in ANN, and the number of hidden layers and neurons in ANN.
2. Cross-validation is utilized to assess the performance of different hyperparameter combinations and select the optimal set of hyperparameters that yield the best performance on unseen data.

By employing these methodologies, we aim to develop accurate and robust models for predicting house prices based on the given features.

Presentation of Experimental Results:

1. The experimental results include the performance metrics obtained for each model on the testing dataset.
2. These metrics include mean squared error (MSE), root mean squared error (RMSE), and R-squared (coefficient of determination), which measure the accuracy and goodness of fit of the models.

Performance Metrics Used for Evaluation:

1. **Mean Squared Error (MSE)**: Measures the average squared difference between the predicted house prices and the actual prices.

2. **Root Mean Squared Error (RMSE):** Represents the square root of MSE and provides a measure of the average deviation of the predicted prices from the actual prices.

3. **R-squared (R2):** Indicates the proportion of the variance in the target variable (house prices) that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating better fit.

Comparison of Different Models and Techniques:

1. The performance of different models (Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, and ANN) is compared based on the aforementioned performance metrics.
2. The models are evaluated in terms of their ability to accurately predict house prices and their robustness to unseen data.

Visualizations to Illustrate Key Findings:

1. Graphs and charts are used to visualize the experimental results and key findings.
2. Bar charts may be employed to compare the performance metrics of different models.
3. Scatter plots can illustrate the relationship between actual and predicted house prices, providing insights into the models' predictive accuracy.
4. Additionally, visualization techniques such as residual plots may be utilized to analyze the distribution of prediction errors and identify any patterns or trends.

By presenting the experimental results along with appropriate visualizations, we aim to provide a comprehensive understanding of the performance of the models and facilitate the comparison of different techniques for house price prediction.

**Discussion:**

Interpretation of the Results and Their Implications:

1. The results obtained from the experimentation provide valuable insights into the performance of various machine learning models for predicting house prices.
2. Interpretation involves understanding the significance of the performance metrics (MSE, RMSE, R-squared) in the context of real estate valuation.
3. Lower values of MSE and RMSE indicate better predictive accuracy, while higher R-squared values suggest a stronger correlation between predicted and actual prices.
4. Implications of the results may include identifying the most effective model for accurate price prediction and understanding the factors that influence house prices the most.

Analysis of the Strengths and Weaknesses of the Models:

1. Each machine learning model has its strengths and weaknesses, which are reflected in their performance metrics.
2. Linear Regression, for example, offers simplicity and interpretability but may struggle with capturing nonlinear relationships in the data.
3. Lasso and Ridge Regression, on the other hand, provide regularization to prevent overfitting but may lead to biased estimates if the regularization parameter is not properly tuned.
4. ANN, being a more complex model, has the potential to capture intricate patterns in the data but requires careful tuning of hyperparameters and is more computationally intensive.

Explanation of Any Unexpected Outcomes or Observations:

1. Unexpected outcomes or observations may include instances where a model performs significantly better or worse than expected.
2. These outcomes could stem from issues such as data quality, feature selection, model complexity, or parameter tuning.
3. Explaining these unexpected outcomes is crucial for understanding the limitations of the models and identifying areas for improvement in future research.

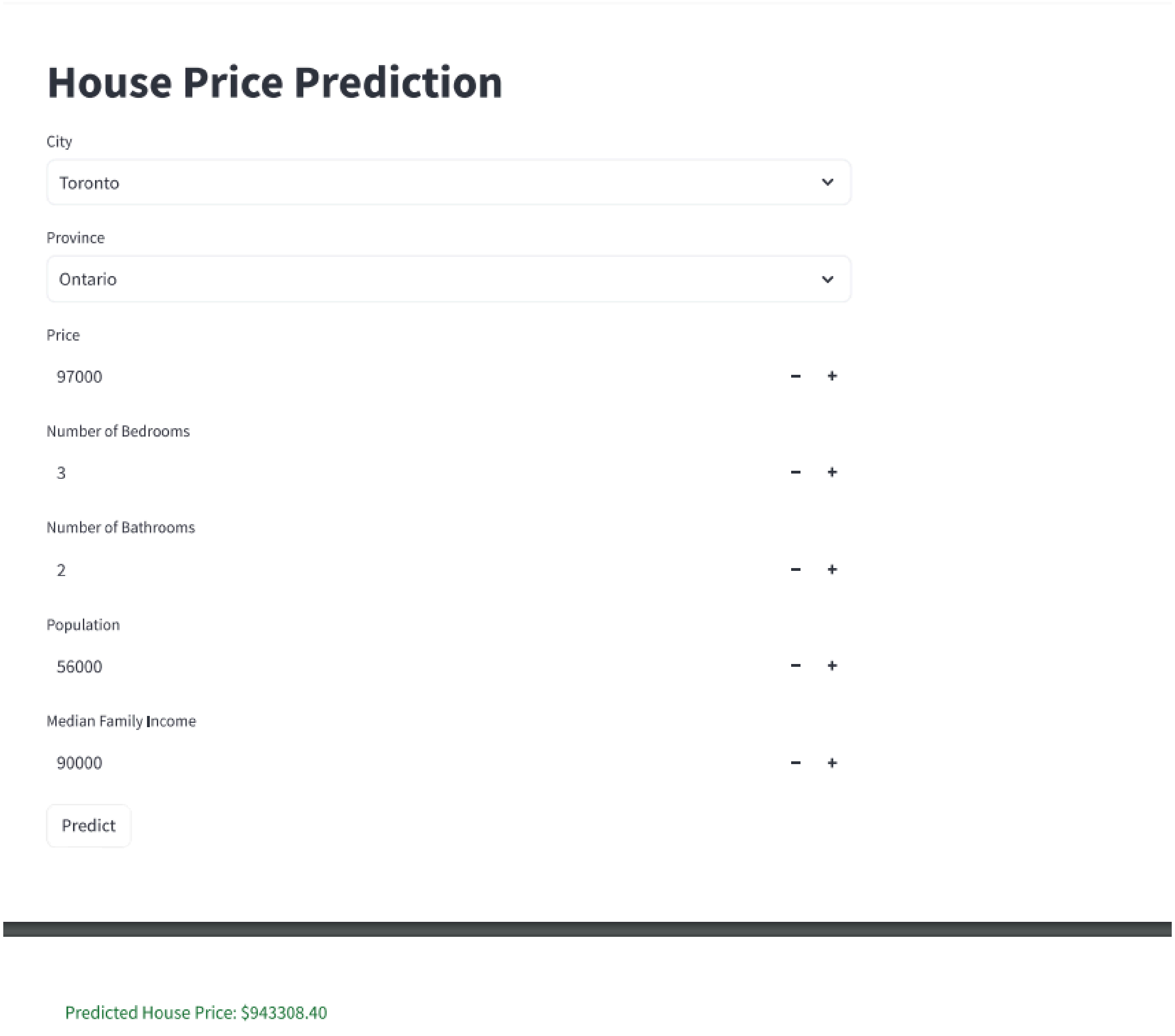
Comparison with Prior Work and Discussion of Contribution to Existing Knowledge:

1. Comparing the project's results with prior research in the field of real estate valuation provides context for evaluating the novelty and significance of the findings.
2. Discussing how the project builds upon or diverges from existing knowledge helps to establish its contribution to the field.
3. This discussion may involve highlighting novel methodologies, addressing limitations of previous studies, or uncovering new insights into the factors influencing house prices.

Overall, the discussion section provides a comprehensive analysis of the experimental results, offering insights into the strengths and limitations of the models, as well as their implications for real-world applications in the domain of real estate valuation.

Model Deployment:

We developed a Streamlit web application for predicting house prices using machine learning models. This application allows users to input information about a property, such as its location, number of bedrooms and bathrooms, population, and median family income. Upon clicking the "Predict" button, the application utilizes a pre-trained machine learning model to generate a predicted house price based on the provided information.



House Price Prediction

City: Toronto

Province: Ontario

Price: 97000

Number of Bedrooms: 3

Number of Bathrooms: 2

Population: 56000

Median Family Income: 90000

Predicted House Price: $943308.40

Key Features:

User-Friendly Interface: The application features an intuitive interface with dropdown menus, number input fields, and a button for prediction.

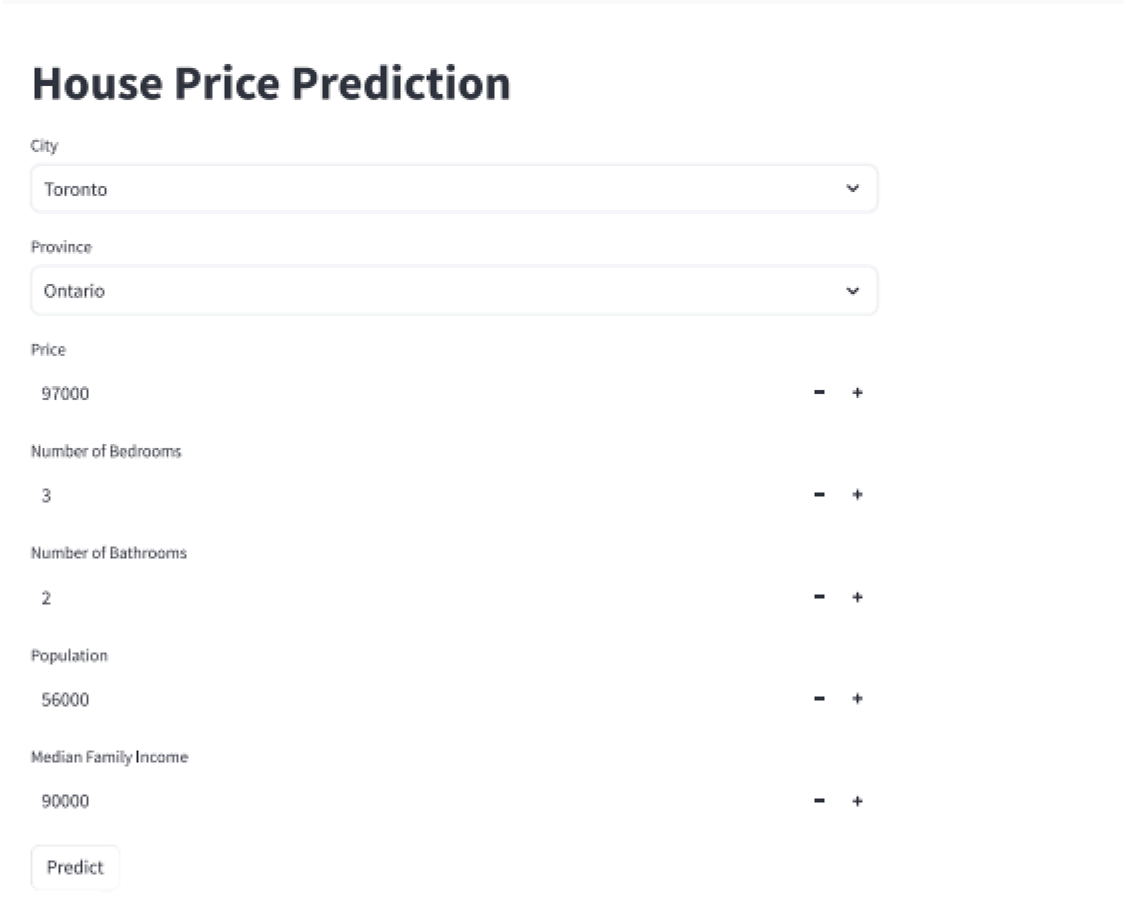
Input Validation: Users can input information such as city, province, price, number of bedrooms and bathrooms, population, and median family income. Input validation ensures that the provided data is within reasonable bounds.

Prediction Display: After clicking the "Predict" button, the application quickly processes the input data, runs it through the machine learning model, and displays the predicted house price in real-time.

Scalability: The application is designed to handle predictions for individual properties efficiently, making it scalable for use with large datasets.

Usage Recommendation: This Streamlit application serves as a valuable tool for real estate professionals, homeowners, and anyone interested in estimating house prices based on various factors. It provides a quick and convenient way to obtain price estimates for properties, aiding in decision-making processes related to buying, selling, or valuing real estate.

Future Enhancements: Potential future enhancements for the application include integrating additional machine learning models for comparison, improving the user interface with additional features and visualizations, and deploying the application to a web server for broader accessibility.



House Price Prediction

City: Toronto

Province: Ontario

Price: 97000

Number of Bedrooms: 3

Number of Bathrooms: 2

Population: 56000

Median Family Income: 90000

**CONCLUSION:**

Summary of Key Findings:

1. The project aimed to predict house prices using various machine learning models and techniques.
2. Experimental results demonstrated the performance of Linear Regression, Lasso Regression, Ridge Regression, Elastic Net, and Artificial Neural Network (ANN) models.
3. Key findings include the identification of the best-performing model based on evaluation metrics such as MSE, RMSE, and R-squared.

Achievement of Project Objectives:

1. The project successfully addressed its objectives by:
2. Collecting and preprocessing real estate data from diverse sources.
3. Implementing machine learning algorithms to train predictive models.
4. Evaluating and comparing model performance using appropriate metrics.
5. Providing insights into the factors influencing house prices and their predictive accuracy.

Recommendations for Future Work or Areas for Improvement:

1. Future research could focus on:
2. Refining feature engineering techniques to capture more nuanced relationships between features and house prices.
3. Exploring advanced machine learning algorithms or ensemble methods to further improve predictive accuracy.
4. Incorporating additional data sources, such as neighborhood characteristics, property amenities, or economic indicators, to enhance model robustness.
5. Conducting longitudinal studies to assess model performance over time and adapt to evolving market trends.

In conclusion, the project contributes valuable insights into real estate valuation using machine learning approaches and lays the groundwork for further research in this domain. By addressing key objectives and highlighting areas for improvement, it serves as a foundation for future endeavors aimed at enhancing the accuracy and applicability of predictive models in the real estate industry.