CSP 571 Data Preparation and Analysis
Project Outline & Proposal

Group Member:

Vaishnavi Mule	A20516627
Neha Ramesh Gawali	A20523722
Atharva Nirali	A20517247
Vikas Reddy - Project Lead	A20539316

Project Proposal:

Research goal

The project aims to develop a predictive model that can accurately estimate housing prices in California. The research goal is to leverage a dataset obtained from Kaggle, specifically the "California Housing Prices" dataset, as the primary source of information. By utilizing techniques from data science and machine learning, along with tools like Microsoft Excel, R, and RStudio, the project seeks to create a robust predictive model that takes various housing-related factors into account, such as location, square footage, and the number of rooms, among others, to predict housing prices. The ultimate objective is to provide a valuable resource for prospective homebuyers, real estate agents, and policymakers to make informed decisions related to housing investments, sales, and market trends.

Research questions

1) What are the key features or attributes within the dataset that most strongly influence house prices in California?

2) Can we build an accurate predictive model for house prices based on historical data?

3) How does the location of a house within California impact its price, and can this be quantified?

4) Are there any temporal trends in California's housing market that affect house prices?

5) What is the overall distribution of house prices in California, and are there any outliers?

6) Can we identify any relationships or correlations between house prices and factors like the number of bedrooms, square footage, or neighborhood demographics?

7) How well does the model perform in terms of predicting house prices, and what is the model's accuracy and error rate?

8) What data preprocessing and feature engineering techniques are necessary to improve the model's accuracy?

9) How can this predictive model be useful for real estate professionals, homebuyers, and sellers in California?

10) Are there specific recommendations or insights the model can provide to make informed real estate decisions?

11) How do the results of this project compare to existing studies, such as the one referenced in the IRJET article?

12) What are the limitations and assumptions of the model, and what further improvements or data sources could enhance its accuracy?

Project Outline:

Data sources and reference data

Dataset:

The US Census Bureau has released Census Data for California, which has 20640 records. The sample dataset contains 10 distinct metrics for each Californian block group, such as population, median income, and median housing price. The median house value attribute of the dataset will be predicted utilizing the various features as independent variables.

- https://www.kaggle.com/datasets/camnugent/california-housing-prices?resource=download&select=housing.csv%29

Reference Data:

Median house prices California districts from the 1990 census

- https://www.kaggle.com/datasets/fedesoriano/california-housing-prices-data-extra-features

The Boston house-price data of Harrison, D. and Rubinfeld, D.L.

- https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data

Real Estate listings (900k+) in the US broken by State and zip code

- https://www.kaggle.com/datasets/ahmedshahriarsakib/usa-real-estate-dataset

Property Prices in United States - Cost of Living

- https://www.kaggle.com/datasets/themrityunjaypathak/property-prices-in-united-states

Dataset description

Data Dictionary – Variable and Description

#	Attribute	Data Type	Description
1	longitude	signed numeric - float	Longitude value for the block in California, USA
2	latitude	numeric - float	Latitude value for the block in California, USA
3	housing_median_age	numeric - int	Median age of the house in the block
4	total_rooms	numeric - int	Count of the total number of rooms (excluding bedrooms) in all houses in the block
5	total_bedrooms	numeric - float	Count of the total number of bedrooms in all houses in the block
6	population	numeric - int	Count of the total number of populations in the block
7	households	numeric - int	Count of the total number of households in the block
8	median_income	numeric - float	Median of the total household income of all the houses in the block
9	ocean_proximity	numeric - Categorical	Type of the landscape of the block [Unique Values : 'NEAR BAY', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', ' Y', '<1H OCEAN', 'INLAND', 'NEAR OCEAN', 'ISLAND']
10	median_house_value	numeric - int	Median of the household prices of all the houses in the block

Dataset Size: 20640 rows x 10 columns

Proposed Methodology:

Data Cleaning:

- Handling Missing Values:
  We handle missing values in the dataset by using techniques like imputation (e.g., filling missing values with the median) or removing rows or columns with missing values.

- Outlier Detection:
  Outlier detection methods can be used to identify and potentially handle outliers in the dataset. Techniques such as the Z-score, IQR, or visualization can be used to detect outliers.

Data Transformation:

- Feature Engineering:
  We perform feature engineering to create new features from the existing ones. This may involve creating ratios, combining features, or transforming variables to capture more meaningful information.

- Scaling/Normalization:
  To make sure that numerical characteristics are on a similar scale, we utilize feature scaling or normalization.

- Categorical Data Encoding:
  The categorical features will be encoded into numerical format. One-hot encoding or label encoding may be used depending on the nature of the data.

Data Pipeline:

- Pipeline Construction:
  We construct a data processing pipeline that includes all the necessary data cleaning and transformation steps. This pipeline will ensure that the data is processed consistently for both training and testing datasets.

- Stratified Sampling:
  Stratified sampling will be used to split the dataset into training and testing sets. This technique helps maintain the distribution of target variables in both sets, reducing potential bias.

Model Selection:

- Approach:
  We use a systematic approach to select the most appropriate model. We will consider linear regression, decision tree regression, and random forest regression models.

- Hyperparameter Tuning:
  For the random forest regression model, we may employ hyperparameter tuning using a grid or random search to find the best combination of hyperparameters that optimize model performance.

- Cross-Validation:
  We will use cross-validation to assess the models' performance and ensure that the chosen model is robust and generalizes well.

Feature Selection Requirements:

- Feature Engineering:
  Feature engineering techniques will be used to create new features that might capture important information.

- Binarization:
  Label binarization will be used to transform categorical features into a format suitable for modeling.

- Scaling:
  Numerical features will be scaled to ensure that the model is not biased toward variables with larger scales.

Regression Approaches:

- The goal is to predict the median house value based on other features.

- Models Used:
  We employ regression models, including *linear regression*, *decision tree regression*, and *random forest regression*, which are well-suited for regression tasks.

- Evaluation Metric:
  We will use the Root Mean Squared Error (RMSE) as the evaluation metric for regression models.

Reference/Baseline Model:

- Baseline Model:
  We establish a baseline model by initially training and evaluating a linear regression model. This model serves as a reference point for comparing the performance of more complex models.

- Model Comparison:
  We compare the performance of decision tree and random forest regression models against the baseline linear regression model to assess whether the more complex models offer significant improvements.

By following these steps and approaches, we aim to select the most appropriate model for our regression task, ensure that the data is properly prepared, and create a reference point for model performance evaluation. This systematic process helps us make informed decisions regarding model selection and hyperparameter tuning.

Literature review:

Investing in real estate is a common desire due to its stability. However, predicting house prices is challenging due to various influencing factors to gain background knowledge of the problem we have used following references:

Shahasane, A., Gosavi, M., Bhagat, A., Mishra, N., & Nerurkar, A. (2023, April 4). *House Price Prediction Using Machine Learning*. www.irjet.net; IRJET. https://www.irjet.net/archives/V10/i4/IRJET-V10I4194.pdf.

The paper is about predicting real estate prices using machine learning algorithms and techniques. It considers various factors that affect house prices, such as area, location, population, size, number of bedrooms and bathrooms, parking space, elevator, etc.The paper uses a dataset from a reputed website to perform data analysis and apply linear regression and sklearn models to increase the accuracy. It also covers data cleaning, outlier removal, feature engineering, dimensionality reduction, gridsearchcv for hyperparameter tuning, k-fold cross-validation, etc.The paper aims to build a house price prediction system with a user-friendly front-end that will help users choose their desired destination and get an idea about the price rates.

Eltanani, S. (2022). Combining Machine Learning models to predict House Prices. www.solent.ac.uk; SOLENT. https://www.solent.ac.uk/documents/degree-shows/isaac-ake-project-scaids.pdf.

This paper states the critical importance of accurate house price valuation in real estate decision-making. It emphasizes the role of predictive models and factors influencing property price changes. Numerical property characteristics and spatial data are discussed, highlighting their impact on price predictions. The paper also emphasizes the development of a software tool to enhance house price prediction algorithms. It suggests that the housing market could benefit from improved mechanisms for projecting house values, with regression methods and machine learning techniques being potential options. The study uses the California House Price Prediction data and employs various regression models, concluding that the Random Forest Regression Model is the most suitable for predicting housing prices.

Software packages, applications, libraries, etc.

1. tidyverse:
   This is a collection of R packages designed for data science.

2. caret:
   This package provides a set of functions that attempt to streamline the process for creating predictive models. It contains tools for data splitting, pre-processing, feature selection, model tuning, and other techniques.

3. randomForest:
   This is an R package that provides functions for building random forest models, a popular machine learning algorithm used for both classification and regression tasks.

4. R Studio