Software Requirements Specification (SRS)

Project Title: House Value Prediction System for California Real Estate Agency

Prepared by: Stefano Chen
Date: April 22, 2025

---

1. Introduction

1.1 Purpose

This document outlines the software requirements for a machine learning (ML) system designed to predict the market value of residential properties in California. Commissioned by a real estate agency, the goal is to enhance the accuracy and efficiency of property valuation processes.

1.2 Scope

The system will provide property value predictions using a supervised ML model trained on historical property data. The system will expose a user-friendly interface and an API to allow agents to input property data and receive valuation predictions.

1.3 Definitions, Acronyms, and Abbreviations

- **ML**: Machine Learning
- **API**: Application Programming Interface
- **MAE**: Mean Absolute Error
- **MAPE**: Mean Absolute Percentage Error
- **RMSE**: Root Mean Square Error
- **R2**: R-Squared

---

2. Overall Description

2.1 Product Perspective

The system is a software solution that integrates with client data sources. It comprises:

- A backend ML engine for prediction
- A RESTful API for data communication
- A web-based frontend for internal users (real estate agents)

2.2 Product Functions

- Accept property feature input and return estimated value.
- Train model on historical data with expert estimations.
- Retrain model with new data.
- Track prediction accuracy and provide performance metrics

2.3 User Characteristics

- **Real estate agents**: Basic computer skills; use the interface to input data and receive predictions
- **ML engineers**: Responsible for model lifecycle management, retraining, and monitoring.

2.4 Constraints

- The predictions must have an average error rate less or equal to 20%.
- Predictions must be returned within 2 seconds.
- Must comply with California data privacy laws.

2.5 Assumptions and Dependencies

- Data provided is representative of the housing market.
- The system relies on stable internet connectivity for API and frontend access

---

3. Specific Requirements

3.1 Functional Requirements

- **FR1**: The system shall be initially trained using the provided dataset.
- **FR2**: The system shall accept property feature inputs via the UI and API.
- **FR3**: The system shall return a predicted value with an associated confidence score.
- **FR4**: The system shall log all prediction inputs and results.
- **FR5**: The system shall support scheduled and event-based model retraining.

3.2 Non-Functional Requirements

- **NFR1**: The system shall maintain a prediction error rate below 20%.
- **NFR2**: The system shall return prediction results within 2 seconds.
- **NFR3**: The system shall be deployed in a secure, scalable cloud environment.
- **NFR4**: The user interface shall be intuitive and accessible to non-technical users.
- **NFR5**: The system shall support version control for both models and data.

3.3 External Interface Requirements

- **User Interface**: A responsive web interface for data input and visualization.
- **API Interface**: RESTful endpoints for system integration and automation workflows.
- **Data Interface**: Input from CSV files and database connectors for model training and logging.

---

4. Appendices

4.1 Initial Dataset Format

Column Name	Description
Median_House_Value	Median house value for household within a block (measured in USD) [prediction target]
Median_Income	Median income for households within a block of houses (measured in tens of thousands of USD) [10k$]
Median_Age	Median age of a house within a block; a lower number is a newer building [years]
Tot_Rooms	Total number of rooms within a block
Tot_Bedrooms	Total number of bedrooms within a block
Population	Total number of people residing within a block
Households	Total number of households, a group of people residing within a home unit, for a block
Latitude	A measure of how far north a house is; a higher value is farther north [°]
Longitude	A measure of how far west a house is; a higher value is farther west [°]
Distance_to_coast	Distance to the nearest coast point [m]
Distance_to_LA	Distance to the centre of Los Angeles [m]
Distance_to_SanDiego	Distance to the centre of San Diego [m]
Distance_to_SanJose	Distance to the centre of San Jose [m]
Distance_to_SanFrancisco	Distance to the centre of San Francisco [m]

4.2 Model Evaluation Metrics

- **MAE (Mean Absolute Error):** Measures average absolute difference between predicted and actual prices

- **RMSE (Root Mean Square Error):** Penalizes larger errors more heavily than **MAE**

- **MAPE (Mean Absolute Percentage Error):** Provides error in percentage terms for business interpretability

- **R2 (R-Squared):** is a statistical measure used to evaluate how well a **regression model** explains the variability of the target variable (dependent variable) based on the input features (independent variables).