Practicum 2 Project – Predicting California Housing Prices Using Deep Learning for Real-World Decision Support

Student: Salman Shareef
Course: Practicum
Instructor: Dr. Follis

Executive Summary

Housing price forecasting is a core component of financial planning for **real estate investors**, **lenders**, **city planners**, and **affordable housing agencies**. Traditional regression models often fail to capture nonlinear interactions between socioeconomic and geographic factors.

This practicum project develops a **Deep Learning–based housing price prediction system** using the California Housing dataset. Four models were evaluated—SVR, KNN, Baseline MLP, and Improved MLP—with the Baseline MLP achieving the highest predictive performance ($\mathbf{R^2 = 0.78}$, RMSE = 0.537).

These results demonstrate that deep learning provides a reliable modeling framework for high-level decision support in:

- Investment risk evaluation
- Housing affordability analysis
- Mortgage underwriting
- Demographic-driven planning

This project proposes not only the model selection but also **how such a system could be deployed and applied in real-world practice**, making it suitable for Practicum 2 standards.

1. Problem Statement (Business & Applied Framing)

The goal is to build a predictive system capable of estimating median house prices across California using census-level features.

Accurate predictions support:

1. Investors

- Identify undervalued regions
- Predict ROI before purchase

2. Banks & Lenders

- Improve mortgage approval risk scoring
- Reduce default risk through price validation

3. City Planners

- Detect high-value or low-value clusters
- Guide resource allocation & zoning decisions

4. Affordable Housing Agencies

- Identify pricing pressure in specific communities

Rather than treating this as a purely academic exercise, this project answers:

“How can a predictive model support real-world decision-making?”

2. Data Description (Applied)

The California Housing dataset contains **20,640 rows × 9 variables**:

- **Economic:** Median income
- **Housing characteristics:** Rooms, bedrooms, occupancy
- **Demographics:** Population
- **Geospatial:** Latitude, longitude
- **Target:** Median house value

Why this dataset is suitable for Practicum work:

- Large enough to require ML pipelines
- Rich real-world characteristics
- Strong socioeconomic relevance
- Clearly applicable to policy, real estate, and finance

3. Exploratory Data Analysis

Key insights:

1. Income is the strongest predictor

Higher median income → substantially higher home value.
This mirrors **real economic behavior** and supports the model's interpretability.

2. Geographic location matters

Latitude/longitude reveal coastal vs. inland price shifts.
This is essential for:

- identifying high-demand ZIP codes,
- planning city infrastructure,
- guiding investors.

3. Housing density varies widely

Population and occupancy show right-skewed distributions.
This signals congestion in certain districts, influencing:

- development planning,
- transportation network design.

4. Multicollinearity checks confirm stable model features

All predictors were retained, improving robustness.

4. Modeling Methodology

The project evaluates **four model families**:

1. Support Vector Regression (SVR)

Nonlinear baseline; good for complex boundaries.

2. k-Nearest Neighbors (KNN)

Simple, nonparametric—but struggles with high-dimensional data.

3. Baseline Deep Learning MLP

Two dense layers; strong default nonlinear learner.

4. Improved Deep Learning MLP

Batch normalization + dropout for stability and generalization.

Business-Grade Workflow:

1. **Train-test split**
2. **Feature scaling**

3. **Cross-validation for classical models**
4. **Early stopping for deep learning**
5. **Model comparison based on RMSE, MAE, R^2**

This mirrors a real-world ML development pipeline.

5. Results and Interpretation for Decision-Making

Model	R^2	RMSE	MAE	Interpretation
Baseline MLP	**0.780**	**0.537**	**0.370**	Best balance of flexibility + generalization
Improved MLP	0.772	0.547	0.387	Too much regularization (slight underfit)
SVR	0.756	0.566	0.371	Good nonlinear performance
KNN	0.680	0.647	0.437	Weak—distance-based methods fail on this geometry

Interpretation:

- With **78% variance explained**, the model is strong enough for **regional planning**, **market forecasting**, and **risk analysis**.
- Error margins are consistent with real housing market uncertainty (±$35–40k).

6. Training Curve Analysis

Baseline MLP Curve

- Converges quickly
- Training and validation loss remain close
- **Indicates excellent generalization**

Improved MLP Curve

- Smoother but slightly higher validation loss
- **Strong stability** but slightly underfits due to dropout

This analysis helps determine:

- model reliability,
- risk of overfitting,
- suitability for deployment.

7. Residual Analysis

Residual scatter:

- Centered around zero
- No strong heteroscedasticity

Histogram:

- Approximately normal distribution

Meaning:

- **Predictions are unbiased and consistent**
- **Model errors are stable**, making it reliable for financial decisions

8. Feature Importance Interpretation (Stakeholder-Centric)

Ranked from highest to lowest:

1. **Median income**

2. **Latitude**
3. **Longitude**
4. **Average rooms / bedrooms**
5. **Population / occupancy**

Implications:

- Income and location dominate home value → helps identify economic pressure zones.
- Room counts matter → useful for builders & developers.
- Occupancy affects affordability → relevant to public policy.

9. Model Deployment Plan

A real-world deployment pipeline would include:

1. Data Ingestion

- Daily or weekly updates from census or real estate APIs

2. Preprocessing Pipeline

- Scaling
- Feature engineering
- Outlier handling

3. Model Hosting

Options:

- AWS SageMaker
- Google Vertex AI
- Flask/FastAPI microservice

4. Prediction Delivery

- Dashboard for city planning
- API for real estate platforms
- Risk scoring module for lenders

5. Monitoring

- Drift detection
- Periodic retraining
- Error tracking

This demonstrates readiness for actual industry use.

10. Limitations & Risk Analysis

Data limitations

- Based on 1990 census; may not represent current patterns
- Lacks crime, school quality, zoning, and neighborhood data

Model limitations

- Deep learning does not explain causal relationships
- Latitude/longitude lose some spatial richness without clustering

Risk implications

- Underestimating homes can mislead investors
- Overestimating can expose lenders to default risk
- Spatial drift may degrade model performance over time

11. Conclusion

This project successfully demonstrates the value of deep learning for real-world housing price prediction. The Baseline MLP provided the strongest accuracy, showing that even moderately sized neural networks capture nonlinear socioeconomic–geographic interactions effectively.

More importantly, the model supports **real-world decision-making**:

- Investors gain insight into undervalued areas
- Lenders reduce loan risk
- City planners identify price trends
- Affordable housing agencies detect market pressure

With additional data sources and deployment infrastructure, this system could serve as a real operational forecasting tool in the real estate industry.