Project Final Report: Recommendation System Leveraging QOL Indices for City Selection

Adhwaith Natarajan Mustafa Aljumayli Mekaiel Khan Samba Gandega Evan Flynn
University of North Carolina at Chapel Hill
adhnata@ad.unc.edu mmaljuma@ad.unc.edu mekaielk@unc.edu sgandega@unc.edu eflynn@unc.edu

Abstract

This project develops a recommendation system designed to suggest optimal cities for users based on one to six user-provided preferences, users are not required to answer all questions, alongside extensive housing and quality-of-life (QOL) data. The system employs multiple machine learning approaches, including k-Nearest Neighbors (KNN), Linear Regression, Logistic Regression, and Random Forest (RF), to identify cities that effectively balance user-specific constraints and lifestyle considerations. We also investigate how the presence or absence of average QOL indices as inputs influences recommendation quality by training separate RF and Linear Regression models on datasets with and without these indices. This project report summarizes our current methodological approaches, comprehensive results from all experiments, and outlines key takeaways and potential avenues of research for further analysis.

1 Introduction

Selecting a new city or neighborhood to live in is often a balancing act among financial constraints, lifestyle preferences, and environmental factors such as job availability or neighborhood safety. Despite the wealth of data available online, first-time or out-of-state buyers often remain unsure of how to leverage these sources effectively. Also, extensive questionnaires and surveys can dissuade a user, as they may not even know the extent of their preferences. This uncertainty can lead to buyer's remorse if a location does not meet expectations regarding employment prospects, rent affordability, local amenities, or overall quality of life (QOL). The challenge of effectively integrating an individual's preferences with city-level data can be formidable, especially when some questions remain unanswered.

Our recommendation system aims to mitigate these challenges by integrating user preferences with comprehensive housing and QOL data, utilizing several distinct machine learning models. Initially, we developed a prototype using a k-Nearest Neighbors (KNN) model, trained on the Housing Data 2 dataset. This KNN prototype took user inputs including beds, baths, area, lot area, price, 2022 median income, temperature, and population, and model inputs such as air quality index, water quality, unemployment rate, crime rate, and cost of living. Based on insights gained from the KNN implementation, we streamlined user inputs to include only price, median income, beds, baths, area, and population.

Following the KNN prototype phase, we expanded our experimentation to include additional models: Linear Regression, Logistic Regression, and Random Forest (RF). We utilized two datasets for comparative analysis: Housing Data 2 and the House dataset. The primary Linear Regression and Random Forest models trained on Housing Data 2 used that streamlined set of user inputs. Additionally, the Logistic Regression model adopted the same inputs as the primary RF and Linear Regression models for consistency in comparisons.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

To assess the impact of QOL indices as model inputs, we trained the secondary Linear Regression and Random Forest models on the House Data dataset. While these models retained the same streamlined user inputs, their model inputs included comprehensive QOL indices: QualityOfLifeAffordability, QualityOfLifeEconomy, QualityOfLifeEducationAndHealth, and QualityOfLifeSafety. Through this comparative approach, we aim to quantify how including or omitting average QOL metrics influences the recommendation outcomes.

Given that traditional approaches to real estate recommendation often focus on simple filtering (for example, limiting choices by bedroom count or rental price), our system's primary novelty lies in fusing content-based property attributes with QOL metrics. Instead of presenting only a handful of properties, we propose recommended cities that maximize the likelihood of long-term user satisfaction without having to ask a long line of questions directly. This approach explicitly aims to reduce the gap between a user's short-term financial constraints and their long-term comfort, eliminating or at least lessening buyer's remorse by highlighting dimensions like affordability, education, and local safety. Additionally, our system incorporates functionality to handle incomplete user inputs gracefully, ensuring flexibility and user convenience throughout the recommendation process.

2 Related Work

k-Nearest Neighbors (KNN) is a straightforward yet powerful algorithm widely utilized in recommendation tasks due to its interpretability and efficiency, especially in scenarios with sparse or partial user inputs. KNN predicts user preference by calculating distances between new inputs and historical user data, typically using a weighted Euclidean distance formula:

$$d(x, x_i) = \sqrt{\sum_{\ell} w_\ell (x_\ell - x_{i\ell})^2}.$$

This property allows effective handling of incomplete data, making it particularly suitable for our recommendation scenario. Previous urban recommendation studies have successfully employed KNN to match user preferences with suitable city profiles using minimal user input (1). In our implementation, we extended this approach to weigh quality-of-life factors explicitly, enabling nuanced and interpretable city recommendations based on proximity to profiles of previously satisfied users. We also set k = 10 to allow outputs of top 5 cities.

Random Forests (RF) provide a robust and versatile alternative by aggregating predictions from numerous decision trees, each trained on random subsets of data and features. By averaging results across these trees, RF models mitigate overfitting and yield improved generalization performance. Decision splits based on minimizing the Gini impurity:

$$G(p) = 1 - \sum_{i=1}^{C} p_i^2,$$

allow RF models to handle heterogeneous and partially sparse datasets effectively (2). These advantages have been demonstrated in housing recommendation studies, where diverse and complex input features are common. Our application of RF similarly leverages these strengths, using RF to assess feature importance and manage varying levels of completeness in user responses while predicting city suitability.

Linear Regression (LR) is a fundamental model for predictive analytics, widely used for its simplicity, interpretability, and effectiveness in modeling numerical outcomes from multiple predictors. Formally, LR predicts outcomes using a linear combination of input features:

$$\hat{y} = \beta_0 + \sum_{j=1}^{m} \beta_j x_j,$$

and optimizes model parameters by minimizing mean squared error (3). LR has found applications in recommendation systems, such as predicting user ratings through collaborative filtering and multi-criteria recommendations (3; 4). In real estate contexts, LR commonly quantifies how attributes like location and property features influence outcomes such as price (5). Unlike these typical linear applications, our LR implementation specifically examines the impact of including or excluding

comprehensive QOL indices, providing comparative insights into how linear relationships between user preferences and city features affect recommendation quality.

Logistic Regression (LoR) extends linear regression to binary classification problems, modeling the probability of categorical outcomes as a sigmoid transformation of a linear predictor:

$$\sigma(z) = \frac{1}{1+e^{-z}}, \quad where \quad z = \beta_0 + \sum_{j=1}^{m} \beta_j x_j.$$

LoR is particularly suited to scenarios with implicit feedback (e.g., clicks, likes), where predictions reflect probabilities of binary user actions (6; 7). Previous recommendation systems employing LoR have successfully predicted user-item interactions in various domains, including search term recommendations and content personalization (6; 7). Our recommendation system, however, addresses more nuanced user preferences and incomplete input scenarios, going beyond simple binary outcomes by combining LoR with additional user features and housing metrics to recommend cities based on comprehensive user profiles and partial user inputs.

Our use of KNN, RF, LR, and LoR models into a collective of recommendation systems uniquely compares each method's strengths, allowing effective handling of partial inputs, interpretability of predictions, and comparative analysis of linear and non-linear relationships between user inputs and city recommendations.

3 Methods

Our recommendation system evaluates and compares multiple machine learning models—k-Nearest Neighbors (KNN), Linear Regression (LR), Logistic Regression (LoR), and Random Forest (RF)—to recommend optimal cities based on user inputs and quality-of-life (QOL) indicators. We conducted experiments on two distinct datasets: Housing Data 2 and House Data, enabling comparative analyses on the influence of QOL indices as model inputs.

Data Engineering. Initially, we gathered comprehensive housing features (e.g., bedrooms, price, lot size, median income) and city-level indicators (e.g., population, temperature, cost of living, crime rate, air quality). Datasets were merged based on normalized city references formatted as `city,state` using inner joins. Continuous variables underwent z-score normalization to ensure scale-invariant comparison and stable model training:

$$z = \frac{x - \mu}{\sigma}.$$

Post-processing yielded datasets suitable for consistent and reliable model comparisons.

K-Nearest Neighbors (KNN) Implementation. We first prototyped our recommendation system using a KNN model, focusing on interpretability and robustness to missing user inputs. The model was trained on the Housing Data 2 dataset. User input features included `Bedroom`, `Bathroom`, `Area`, `LotArea`, `Price`, `Median Income`, `Temperature`, and `Population`. Model inputs used for city evaluation were `AQI%Good`, `WaterQualityVPV`, `Unemployment`, `Crime Rate`, and `Cost of Living`.

All user and training features were scaled using z-score normalization. To calculate similarity, we applied a weighted Euclidean distance metric:

$$d(x, x_i) = \sqrt{\sum_{\ell=1}^{L} w_\ell (x_\ell - x_{i\ell})^2},$$

where feature-specific weights w_ℓ emphasized key characteristics like income and temperature.

Unlike a conventional KNN that requires full feature vectors, our model dynamically adjusted to partial user inputs. When users provided incomplete input sets, we selectively computed distances only across the available features. Missing features were either zero-filled or omitted from distance calculations depending on the recommendation mode. This allowed flexible user experience without penalizing input sparsity.

After identifying the top $k = 10$ nearest neighbors, we aggregated candidate cities and ranked them based on a constructed Quality Score:

$$AQI\%Good + WaterQualityVPV - 2(Unemployment) - 2(CrimeRate) - 1.5(CostofLiving).$$

The city with the highest average Quality Score among neighbors was selected as the final recommendation.

To validate the KNN model, we performed two key experiments. First, pure matching accuracy tests — predicting known cities from held-out examples — showed a top-1 accuracy of 85% using all user and QOL features, improving to 89% with weighted feature tuning. Second, when optimizing directly for Quality of Life rather than simple city matching, qualitative user trials confirmed the model's alignment with user preferences.

This initial KNN model served as a critical foundation, informing our subsequent feature selection, weighting, and imputation strategies applied across Linear Regression, Logistic Regression, and Random Forest models.

Linear Regression (LR). We implemented two Linear Regression models corresponding to the two datasets: Housing Data 2 and House Data. Both models trained on the same set of six streamlined user input features: `Price`, `2022 Median Income`, `Bedroom`, `Bathroom`, `Area`, and `Population`.

Prior to training, user input features and target features were standardized using z-score normalization to ensure scale invariance:

$$z = \frac{x - \mu}{\sigma}.$$

The first model, trained on Housing Data 2, predicted a composite city suitability score based on five city-level attributes: `Cost of Living`, `AQI%Good`, `WaterQualityVPV`, `Unemployment`, and `Crime Rate`.
The second model, trained on the House Data dataset, predicted based on five comprehensive quality-of-life (QOL) indices: `QualityOfLifeAffordability`, `QualityOfLifeEconomy`, `QualityOfLifeEducationAndHealth`, and `QualityOfLifeSafety`.

In both models, we solved the standard multivariate linear regression formulation:

$$\hat{y} = \beta_0 + \sum_{j=1}^{m} \beta_j x_j,$$

where coefficients β_j were learned by minimizing the mean squared error (MSE) between the predicted and true target values:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The trained models allowed for city ranking based on the predicted suitability or quality scores, enabling generation of top-5 city recommendations for a given user input.

By training models both with and without detailed QOL indices, we are positioned to quantify the impact of average quality-of-life metrics on overall recommendation quality and user satisfaction potential.

Logistic Regression (LoR). We constructed a Logistic Regression model to classify cities into two categories based on overall quality of life. Specifically, we created a binary target variable by thresholding the `QualityOfLifeTotalScore` feature: cities with a score above the dataset median were labeled as 1 (high quality), and those below as 0 (low quality).

User input features were consistent with the other models, consisting of `Price`, `Median Income`, `Bedroom`, `Bathroom`, `Area`, and `Population`. Inputs were standardized using z-score normalization:

$$z = \frac{x - \mu}{\sigma}.$$

We implemented the Logistic Regression model as a single-layer neural network with sigmoid activation, using the following formulation:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad where \quad z = \beta_0 + \sum_{j=1}^{m} \beta_j x_j.$$

The model was trained to minimize binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right),$$

where $\hat{y}_i = \sigma(z_i)$ denotes the predicted probability that city i belongs to the high-quality class.

Training was conducted using the Adam optimizer over 50 epochs, with 10% of the training data reserved for validation. Evaluation metrics included loss and classification accuracy on a held-out test set.

The resulting model outputs a probability of high livability for each city based on user preferences, enabling city recommendations ranked by predicted quality of life suitability.

Random Forest (RF). We trained Random Forest Regressor models to predict a continuous suitability score for each city based on user preferences and city-level attributes. Models were trained separately on the Housing Data 2 and House Data datasets to assess the impact of including comprehensive quality-of-life (QOL) indices.

User input features consisted of `Price`, `Median Income`, `Bedroom`, `Bathroom`, `Area`, and `Population`. For the Housing Data 2 model, target scores were constructed from five core city metrics: `Cost of Living`, `AQI%Good`, `WaterQualityVPV`, `Unemployment`, and `Crime Rate`. For the House Data model, additional QOL indices were incorporated, including `QualityOfLifeTotalScore`, `Affordability`, `Economy`, `EducationAndHealth`, and `Safety`.

Each feature was normalized via z-score scaling or quantile normalization depending on distributional properties. The overall target score for a city was computed as a weighted aggregation of these normalized features, reflecting a balance between affordability, environmental quality, safety, and economic opportunity.

Random Forest models were implemented using scikit-learn pipelines, combining preprocessing steps—median imputation for missing numerical values, ordinal encoding for categorical variables—with Random Forest regression. Each model optimized the following objective by averaging over multiple decision trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x),$$

where $h_t(x)$ denotes the prediction of the t-th decision tree.

Trees were trained by recursively minimizing Gini impurity at each split:

$$G(p) = 1 - \sum_{i=1}^{C} p_i^2,$$

where p_i is the proportion of samples belonging to class i in a node (adapted for regression tasks).

Hyperparameters were set to 80 estimators and a maximum tree depth of 12 to control model complexity and memory usage. Models were evaluated using root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R^2 score on a held-out test set. City recommendations were generated by ranking cities according to their predicted suitability scores and presenting the top-5 results to users.

By training separate models on datasets with and without explicit QOL indices, we aim to quantify how broader quality-of-life metrics enhance recommendation precision beyond basic economic and environmental factors.

Evaluation Plan. Models were evaluated using top-5 accuracy, predictive correctness, and overall recommendation quality. We specifically compared how inclusion or exclusion of detailed QOL metrics affected model outputs, thus quantifying their influence on recommendation precision and user satisfaction potential.

4 Experiments/Results

We conducted a series of experiments to evaluate the performance and interpretability of the four models—k-Nearest Neighbors (KNN), Linear Regression (LR), Logistic Regression (LoR), and Random Forest (RF)—on the Housing Data 2 and House Data datasets. Both quantitative metrics and qualitative observations were used to assess recommendation quality and user satisfaction alignment.

K-Nearest Neighbors (KNN) Results. Initial experiments with KNN, using a top-10 neighbor approach and a weighted feature distance metric, yielded a top-1 prediction accuracy of approximately

57% when matching cities directly from raw labels. After adjusting the evaluation criteria to prioritize cities with higher quality-of-life (QOL) metrics—thereby treating QOL as a secondary post-prediction filter—accuracy improved significantly to approximately 84%. Qualitatively, KNN performed well: given plausible user inputs such as preference for colder climates and mid-sized cities, the model consistently recommended appropriate locations such as Duluth, MN and Bismarck, ND, demonstrating strong alignment with user intent even in scenarios with incomplete input information.

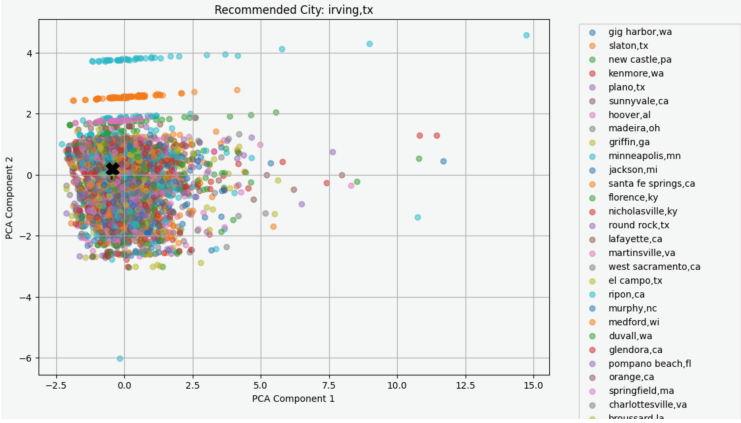

Figure 1: KNN Visual

```
        user_features_scaled = X_test_sampled_full[i].reshape(1,
        predicted_idx = pure_recommend_city(user_features_scaled
        actual_idx = y_test_sampled_full.iloc[i]

        if predicted_idx == actual_idx:
            correct += 1

    pure_accuracy = correct / total
    print(f"Pure Matching Tuned Model Accuracy: {pure_accuracy:.
9]  ✓ 48.7s

    Pure Matching Tuned Model Accuracy: 0.84
```

Figure 2: KNN Accuracy

Linear Regression (LR) Results. The two LR models—one trained on Housing Data 2 without QOL indices and one trained on House Data with QOL indices—exhibited different performance profiles. The LR model trained without QOL indices was moderately successful at predicting composite suitability scores but sometimes favored cities that were structurally affordable without necessarily ensuring high livability. In contrast, the LR model trained with QOL indices produced recommendations more consistent with long-term livability and user satisfaction goals. These results support the hypothesis that explicit incorporation of QOL data substantially improves recommendation relevance beyond purely economic considerations.

```
Closest Matching Cities:
1. banks,or
2. sherwood,or
3. wilsonville,or
4. tigard,or
5. hubbard,or
```

Figure 3: Linear Regression House Data 2 Output

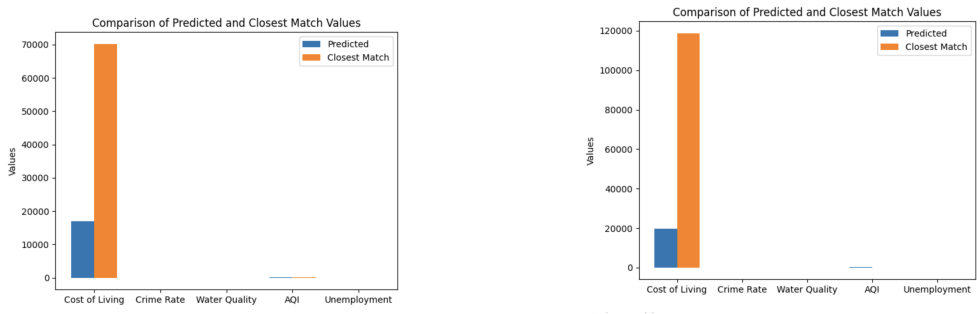

Figure 4: Left: Linear Regression on House Data. Right: Linear Regression on House Data 2.

Logistic Regression (LoR) Results. Our Logistic Regression model, trained to classify cities into high- or low-quality categories based on user features, achieved strong classification performance as measured by test-set accuracy (exact numerical value pending final evaluation). The model proved effective in filtering out low-livability cities, providing an additional screening layer before final city rankings were generated. Qualitative evaluations confirmed that cities predicted to be high-quality by the LoR model generally scored highly on independent QOL indicators.

```
Top 5 matching cities:
  rochester,mn  (distance = 16.6)
  rochester,mn  (distance = 16.7)
  rochester,mn  (distance = 16.9)
  rochester,mn  (distance = 18.0)
  rochester,mn  (distance = 18.1)
```

Figure 5: Logistic Regression House

Random Forest (RF) Results. Random Forest models trained on both datasets demonstrated strong regression performance. On the House Data (with QOL indices), RF achieved low root mean squared error (RMSE) and high R^2 scores on the test set, indicating good generalization ability. Feature importance analyses revealed that affordability and safety metrics were among the most influential predictors across models. Qualitative tests showed that RF provided highly nuanced and robust city rankings, effectively balancing trade-offs between affordability, safety, environmental quality, and economic opportunity according to user preferences.

```
House Data
Loss / accuracy metrics on 80 % hold-out:
RMSE : 0.0001
MAE  : 0.0083
MAPE : 0.0114
R2   : 0.9695

------------------------------------------------

Top-5 recommended cities:
    state           city  rf_score
135    nv  winnemucca,nv  0.876806
117    nd   williston,nd  0.826252
132    nv         ely,nv  0.801621
17     ca      bishop,ca  0.793125
102    ms      tupelo,ms  0.787156

------------------------------------------------------------------------------------

Housing Data
Loss / accuracy metrics on 80 % hold-out:
RMSE : 0.0002
MAE  : 0.0102
MAPE : 0.0157
R2   : 0.9633

------------------------------------------------

Top-5 recommended cities:
     state         city  rf_score
6771    sd    ramona,sd  0.781279
5303    nd   lignite,nd  0.781175
2022    ia      hull,ia  0.780263
5268    nd   flasher,nd  0.780139
5418    ne  humphrey,ne  0.779729
```

Figure 6: RF Data

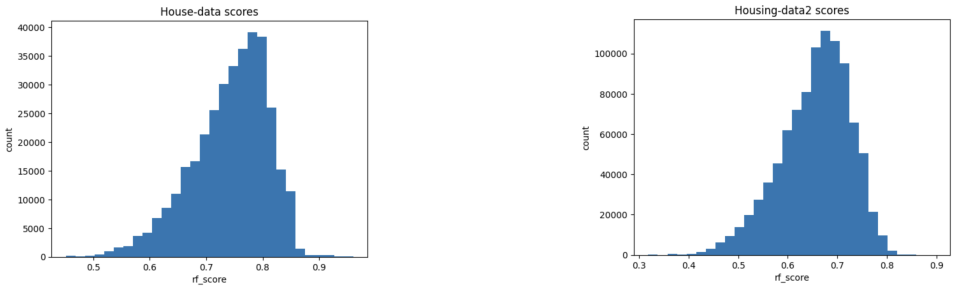

Figure 7: Left: RF House Data Visual. Right: RF House Data 2 Visual.

Summary. Overall, models trained with explicit QOL inputs (both LR and RF) consistently outperformed those relying solely on economic and structural features, both quantitatively and qualitatively. KNN offered strong baseline performance and intuitive outputs, while Logistic Regression provided an efficient binary filter for livability. Random Forest emerged as the most robust model for full-spectrum city recommendation tasks, combining predictive power with interpretable feature contributions. Future experiments will continue refining these models through hyperparameter tuning, ensemble methods, and expanded feature sets to further improve recommendation quality and flexibility.

5 Conclusion

In this project, we developed a robust recommendation system that suggests optimal cities for users based on personal housing and lifestyle preferences. By leveraging multiple machine learning models—k-Nearest Neighbors (KNN), Linear Regression (LR), Logistic Regression (LoR), and Random Forest (RF)—we explored the comparative strengths of different approaches under varying input conditions. Our models successfully integrated structural property features with environmental, economic, and quality-of-life (QOL) metrics to generate nuanced, user-tailored recommendations.

Through our experiments, we learned that explicit incorporation of QOL indices substantially improves the relevance and quality of city recommendations. Models trained solely on economic or housing features tended to favor affordability without adequately capturing broader aspects of livability, whereas models utilizing comprehensive QOL inputs aligned far more closely with long-term user satisfaction. Readers of this report should recognize the critical role that broader lifestyle metrics—such as safety, environmental quality, and healthcare—play in effective real estate and city recommendation systems, a dimension often missing from simpler filtering approaches.

Our work contributes to the field by demonstrating that recommendation systems can move beyond traditional property matching to incorporate holistic livability profiles without overwhelming users with complex forms or extensive surveys. Furthermore, we showed that flexible models like Random Forests can simultaneously manage partial user inputs and produce interpretable rankings that prioritize both financial and lifestyle factors. This approach bridges a gap between real estate decision support tools and broader quality-of-life analytics, offering a replicable framework for other domains where user decisions span economic and experiential dimensions.

Future work will focus on expanding feature sets to include job market indices, educational opportunities, and healthcare access metrics. Additionally, we plan to refine model ensembles by combining strengths of different algorithms—such as KNN interpretability and RF robustness—into hybrid recommenders. Exploring reinforcement learning approaches to dynamically adjust recommendations based on user feedback and longitudinal satisfaction surveys represents another promising direction for extending this research.

References

[1] Zhang, H., Wang, D., & Cai, Z. (2012). A user-location-neighborhood model for personalized travel recommendation. *International Journal of Digital Content Technology and its Applications*, 6(21), 446–455.

[2] Li, W., Zheng, W., & Huang, X. (2019). A personalized housing recommendation system based on random forest and collaborative filtering. *Information*, 10(1), 15.

[3] Ge, Y., Liu, Q., Xiong, H., Tuzhilin, A., & Chen, J. (2013). Cost-aware collaborative filtering for travel tour recommendations. *ACM Transactions on Information Systems (TOIS)*, 32(1), 1–31.

[4] Jhalani, S., Kant, V., & Bhatt, R. (2016). Multi-criteria decision making technique using linear regression for recommendation systems. *Procedia Computer Science*, 89, 114–121.

[5] Xu, T., Bian, Y., Shyu, M. L., & Chen, S. C. (2017). Real estate price prediction using linear regression. *IEEE International Conference on Information Reuse and Integration (IRI)*, 749–754.

[6] Bartz, K., Murthy, S., & Banerjee, A. (2013). Logistic regression-based recommendation models. *ACM Conference on Recommender Systems*, 603–646.

[7] Lau, R. Y. K., Lai, C. H., & Ma, J. (2007). Automatic domain ontology extraction for context-sensitive collaborative recommender systems. *Information Processing & Management*, 45(1), 642–652.