December 12, 2024

Carly Carroll and
Ryan Tsang

Predicting Housing Affordability and Migration Patterns

CSC/DSP 461: Machine Learning

Introduction

Context

- Housing affordability and migration patterns are crucial for the well being of society
- Using machine learning techniques to uncover actionable insights

Objective

- Prediction of housing affordability
- Analyze migration patterns
- Enhance decision making with interactive tools

Significance

- Actionable insights for individuals, families, urban planners, and policymakers
- Allowance to build on a foundation that combines affordability and migration to explore new trends and refine models

Problem Definition

Introduction to the Problem

- Affordability and migration patterns are deeply interconnected, with affordability shaping where people live and migration reflecting broader social and economic shifts
- Investigate how housing affordability influences migration patterns by analyzing inflow and outflow data
- Develop an affordability score for each US county that includes housing costs, income levels, local demographics, and living expenses

Objectives of the Study

- Analyze inflow, outflow, and net migration to understand population shifts
- Group counties by shared characteristics to identify regional patterns
- Create models to determine if affordability and migration trends can be predicted for each county
- Provide actionable insights using interactive tools

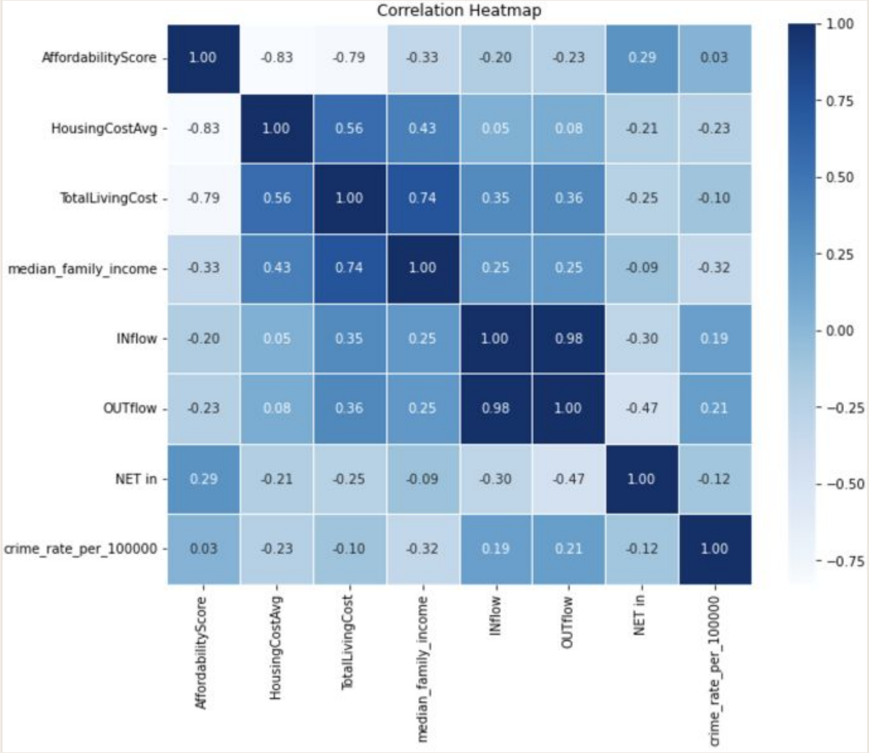
Data Overview

Dataset Characteristics

- Incorporates a variety of socioeconomic, demographic, and migration related characteristics for multiple counties across the United States
- Includes ~2000 rows and a variety of detailed columns

Key Features

- **Affordability:** HousingCostAvg, TotalLivingCost, AffordabilityScore
- **Migration:** INflow, OUTflow, NET in
- **Demographics:** median_family_income, crime_rate_per_100000



Methods Overview

Exploratory Data Analysis

- Gain initial insights into the dataset and understand relationships between key variables
- Calculate descriptive statistics to summarize the distributions of various variables
- Graphical techniques were used to visualize distributions and correlations between variables

Machine Learning

- Clustering techniques were used to group counties based on affordability related characteristics
- Regression models were used to help predict affordability scores
- Migration patterns were predicted using a variety of classification techniques to categorize counties

Neural Networks and Interactive Tools

- Neural networks were used to predict housing affordability scores and classify migration patterns
- K-fold cross validation was applied to both regression and classification methods to ensure robustness
- Interactive tools were created to enhance the accessibility and usability of the predictive models

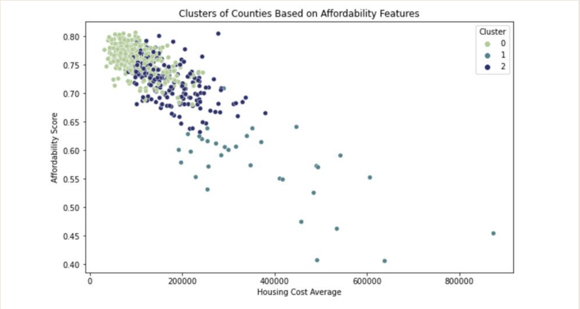
Experiment 1: Data Clustering and Pattern Analysis

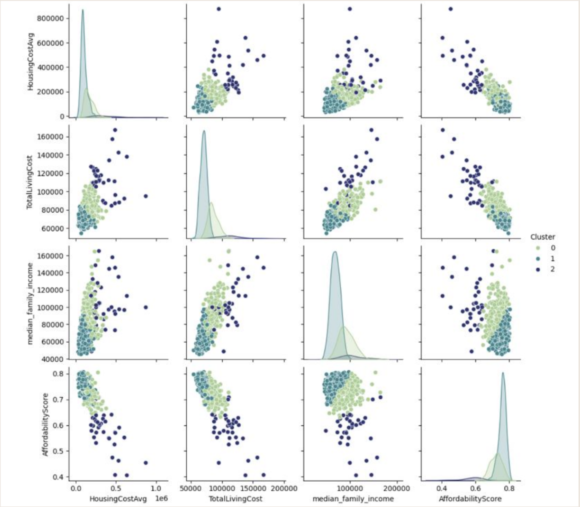
Clustering Methodology

Objective: Group Counties by affordability

Techniques:

- K-means clustering
  - Identifies groups of counties based on affordability
  - Elbow method used to determine optimal number of clusters
- Hierarchical Clustering
  - Validates K-means with dendrogram analysis
  - Focused on top and bottom 10 counties for better interpretation
- Silhouette Analysis:
  - Assess the quality of the clusters
  - Indicates how separated the clusters are
- Classification of Cluster Membership
  - Used a random forest classifier to analyze predictability

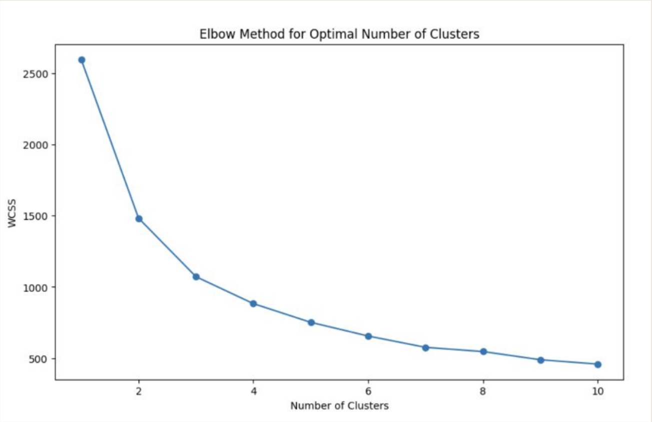


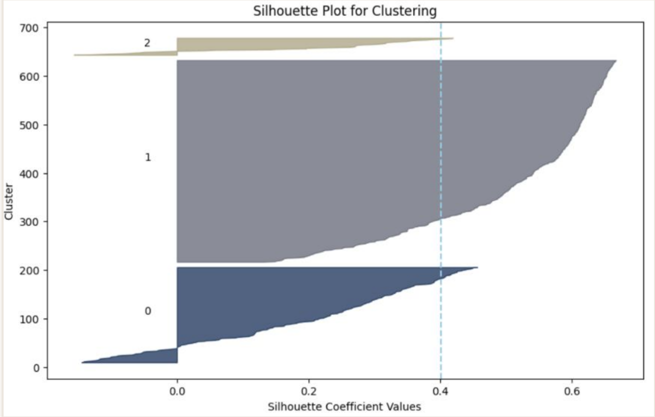


Experiment 1: Data Clustering and Pattern Analysis

Results:

- The elbow method determined the optimal number of clusters to be 3
- The average silhouette score for the 3 clusters was 0.41
  - This score indicates moderate separation
- The two dendrograms displays well defined groupings for the top and bottom 10 counties
- The classification report for predicting cluster membership achieved an impressive accuracy of 97%
  - This accuracy emphasizes the separability of each of the clusters
- The variable TotalLivingCost was identified as the most crucial with an importance of about 40%





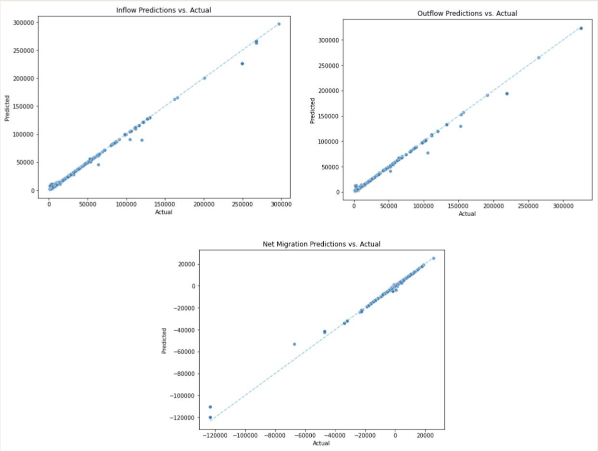
Experiment 2: Predictive Modeling for Trend Analysis

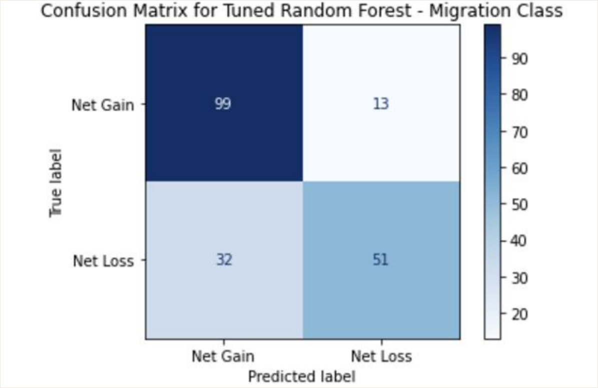
Predictive Modeling Methodology

Objective: Predict migration metrics: INflow, OUTflow, and NET in

Techniques:

- Regression for Migration Metrics
  - Aimed to predict each of the migration metrics
- Feature Importance
  - Plots to determine the most significant predictors
- Classification
  - Categorize the counties into Net Gain and Net Loss
  - Used grid search to identify optimal parameters
- Confusion Matrix
  - Displays the classification model's performance
- Bagging and Boosting
  - Used to compare results with the classification model
- Cross Validation
  - Provided a validation of the predictive models used

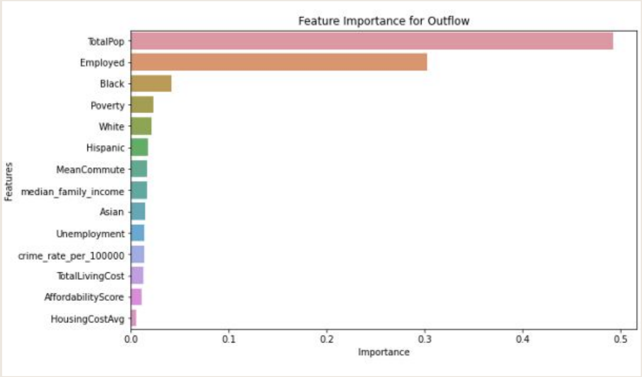


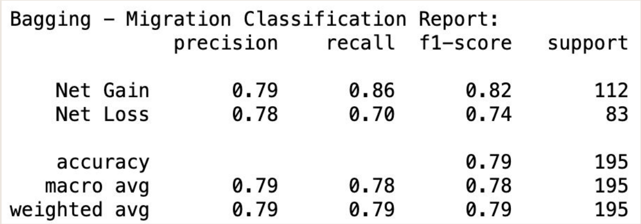


Experiment 2: Predictive Modeling for Trend Analysis

Results:

- The regression model displayed impressive and consistently high R² scores
- Feature importance determined that TotalPop, Employment, and TotalLivingCost were the most significant predictors of Net Migration
- The tuned random forest classification model achieved an accuracy of 77% on the test data
  - Showed a strong performance when identifying Net Gain counties while Net Loss was less robust
- Confusion matrix showed a higher rate of misclassifications for Net Loss (32 instances misclassified as Net Gain)
- Bagging (79%) slightly outperformed boosting methods (69%), exhibiting similar trends to that of the confusion matrix and tuned random forest
- Cross validation achieved an MAE of 0.098 indicating that the regression models maintained low error rates

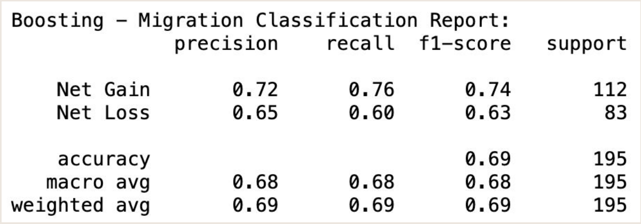




```
Bagging - Migration Classification Report:
              precision    recall  f1-score   support

    Net Gain       0.79      0.86      0.82       112
    Net Loss       0.78      0.70      0.74        83

    accuracy                           0.79       195
   macro avg       0.79      0.78      0.78       195
weighted avg       0.79      0.79      0.79       195
```



```
Boosting - Migration Classification Report:
              precision    recall  f1-score   support

    Net Gain       0.72      0.76      0.74       112
    Net Loss       0.65      0.60      0.63        83

    accuracy                           0.69       195
   macro avg       0.68      0.68      0.68       195
weighted avg       0.69      0.69      0.69       195
```

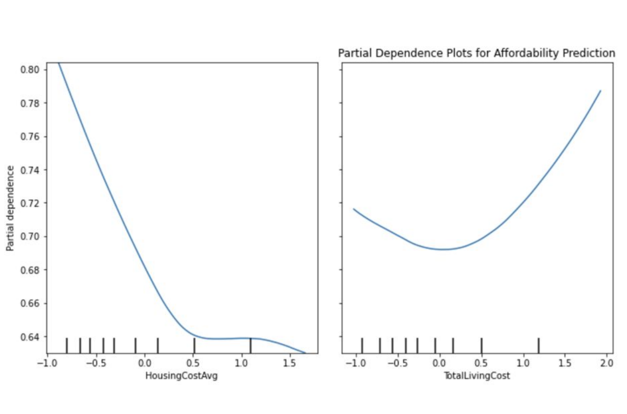
Experiment 3: Neural Network for Predictive Modeling

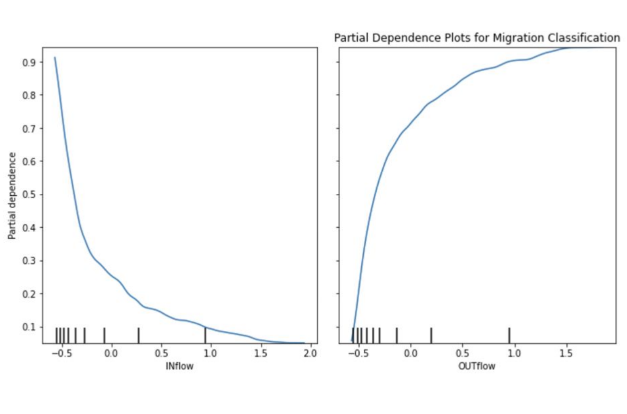
Neural Network Methodology

Objective: Model affordability scores and migration patterns

Techniques:

- Affordability Regression
  - Aimed to predict the affordability score of each county based on socioeconomic and demographic variables
- Classification
  - Categorize counties into Net Gain, Net Loss, and Neutral using migration and socioeconomic variables
- Confusion Matrix
  - Explains the model’s performance displaying the true and predicted values of Net Loss and Net gain
- Feature Importance
  - Determine the top influential predictors of migration class
- Cross Validation
  - Ensures robustness and generalizability of the models

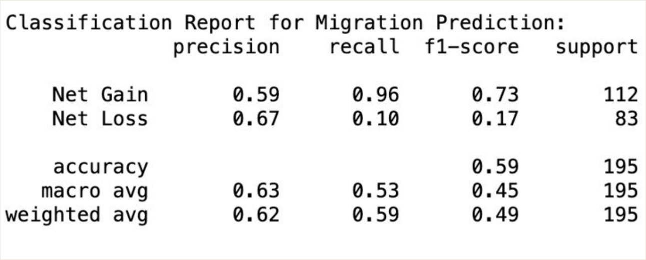




Experiment 3: Neural Network Analysis for Predictive Analysis

Results:

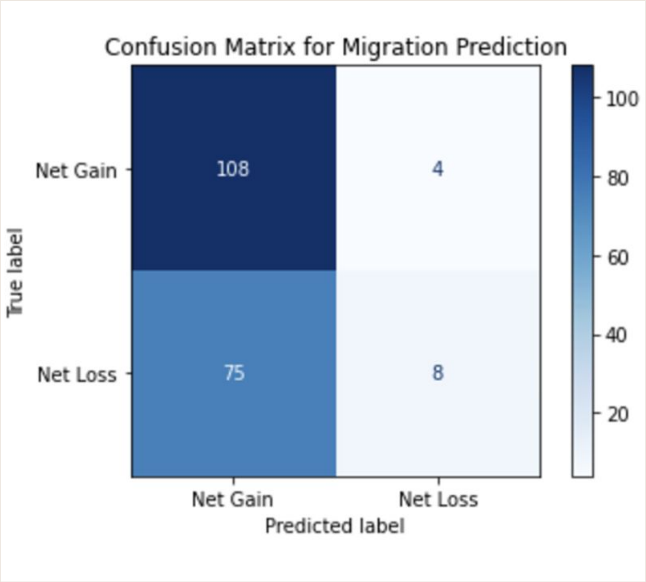
- When performed, the affordability regression model performed poorly
  - Despite having a sufficient MAE of 0.14, the model’s low R² score indicated poor generalization
- Classification for migration patterns achieved an overall accuracy of 59%
  - The Net Gain class displayed much stronger precision compared to Net Loss
- The confusion matrix provided a new view of the model’s performance
  - It displayed similar results to the classification model
- Partial dependency plots revealed non-linear relationships indicating increase housing often correlates with a shift towards Net Loss classification
- For regression, the cross validation had an accuracy of 89.2% demonstrated strong generalizability



```
Classification Report for Migration Prediction:
              precision    recall  f1-score   support

    Net Gain       0.59      0.96      0.73       112
    Net Loss       0.67      0.10      0.17        83

    accuracy                           0.59       195
   macro avg       0.63      0.53      0.45       195
weighted avg       0.62      0.59      0.49       195
```

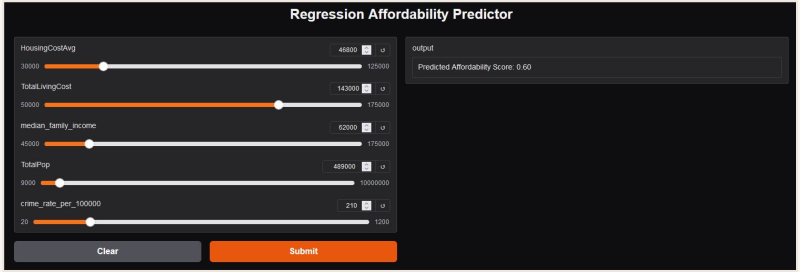


Interactive Tools for Analysis

Objective: Provide a favorable interface for predicting affordability scores and classifying migration patterns

Techniques:

- Gradio for user interfaces
  - User friendly input forms for affordability and migration predictions
  - Outputs predicted affordability score and migration categories
    - Affordability Score: 0-1.0
    - Migration Classification: Net Gain, Net Loss, Neutral
  - Sliders included for controlled inputs
  - Instant prediction feedback displayed in real time
- https://huggingface.co/spaces/22tsangr/demo



Regression Affordability Predictor

HousingCostAvg 46800 (30000 – 125000)

TotalLivingCost 143000 (50000 – 175000)

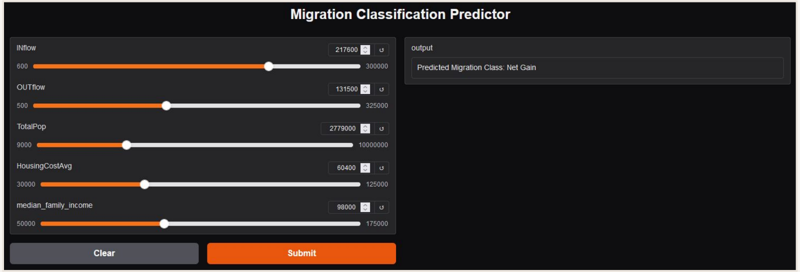
median_family_income 62000 (45000 – 175000)

TotalPop 489000 (9000 – 10000000)

crime_rate_per_100000 210 (20 – 1200)

output

Predicted Affordability Score: 0.60



Migration Classification Predictor

INflow 217600 (600 – 300000)

OUTflow 131500 (500 – 325000)

TotalPop 2779000 (9000 – 10000000)

HousingCostAvg 60400 (30000 – 125000)

median_family_income 98000 (50000 – 175000)

output

Predicted Migration Class: Net Gain

Summary of Key Findings

Experiment 1

- Identified three distinct clusters of counties based on affordability and migration patterns
    - Cluster 0: Moderate affordability and balanced living costs
    - Cluster 1: Economically stable counties with low housing costs
    - Cluster 2: High cost counties with significant outflows
- Key features influencing cluster classification include HousingCostAvg, TotalLivingCost, and median_family_income

Experiment 2

- Achieved high accuracy in predicting inflow, outflow, and net migration, demonstrating their effectiveness in capturing relationships between predictors and migration trends
- Random forest classification reach an accuracy of 77% highlighting migration flow and income as influential predictors
- Effectively identified counties experiencing Net Gain but was challenged by those experiencing Net Loss

Summary of Key Findings

Experiment 3

- Regression neural networks achieved a cross validation MAE 0f 0.098 indicative of robust performance across data subsets
- Classification neural networks excelled in predicting Net Gain categories but struggled with Net Loss predictions
- Partial dependency plots provided insights into feature interactions displaying that the rise in housing and living costs negatively impact affordability and migration outcomes

Interactive Tools

- Developed Gradio based real time interfaces for affordability and migration predictions
- Provides an user friendly and accessible way for user to engage with models and explore predictions interactively

Key Implications

Policy Impact

- Identifying counties with extreme affordability and migration patterns enables policymakers to allocate resources effectively
- Insights into migration trends can guide economic motivation for underdeveloped regions

Urban Planning

- Highlighting affordability hotspots supports sustainable urban and regional development
- Anticipating migration patterns supports infrastructure planning and community service optimization

Economic Insights

- Understanding the relationships between affordability and migration promotes regional economic stability
- Models provide actionable insights to combat housing crises and support growth

Accessibility

- Interactive tools allow for access to predictive models in a user friendly manner
- Enhanced decision making through transparent and interpretable results

Limitations and Future Directions

Limitations

- **Data Imbalance:** While all US States (excluding territories) were included, not all state counties were accounted for, leading to some states having more data to work with than others
- **Neural Network Discrepancies:** Differing test and cross validation metrics suggest further refinement is needed
- **Input Limits:** Going outside of the limits of the training data can cause poor outputs, so the regression and classification models are bounded to around the minimum and maximum of the training data

Future Directions

- **Alternative Model Architecture:** Exploring different libraries such as XGBoost or RMSProp could yield better results
- **Expansion of Dataset:** Sifting through more sources or the addition of temporal data could reveal more patterns that could be used to more accurately make predictions
- **Expansion of Interactive Tools:** Features such as visualization of a datapoint to the regression line or an output showing how classification was assigned could further improve the understanding of these models