Exploring Factors Influencing Homeownership Using Support Vector Models

Name: Aarushi Kotwani

ABSTRACT

This report explores the use of support vector models for classification, in predicting whether a dwelling is occupied by owners or renters based on demographic and housing variables. The data, obtained from the US Census via IPUMS (1) USA, contains information about individuals and their housing situations, including age, income, education level, and various housing characteristics. The analysis is initiated by narrowing down the dataset to include only married individuals whose spouses are present. The report employs support vector machines (SVMs) with linear, radial, and polynomial kernels to model the relationship between predictor variables and homeownership status. The SVM models achieved promising results, with accuracy rates of 86.40% for linear SVM with C=1, 87.65% for radial SVM with C=1 and gamma=10, and 88.11% for polynomial SVM with C=1 and degree=4. Key findings suggest that age, income, and home size significantly impact homeownership status. These factors emerged consistently across different kernel types, underlining their importance in predicting homeownership.

INTRODUCTION

In this analysis, exploration revolves around the dynamics of homeownership in Washington State. The primary objective is to uncover the influences shaping whether a dwelling is occupied by homeowners or renters. To achieve this, a tool called support vector machines is utilized to analyze data collected by the census.

The dataset used in this analysis contains a wide range of variables related to both individuals and housing units. Variables include demographic information such as age, income, and details about their homes, including construction dates and utility costs. By studying this data and employing support vector machines, the aim is to identify which factors hold the most significance in determining whether a dwelling is occupied by homeowners or renters.

However, the goal extends beyond mere prediction. Understanding the rationale behind these predictions is essential. What factors drive individuals to become homeowners while others opt to rent? By addressing this question, the aspiration is to furnish policymakers with valuable insights, ensuring equitable access to homeownership for all in Washington State.

THEORITICAL BACKGROUND

The Support Vector Machine (SVM) is a machine learning technique used to classify data into distinct categories. Its main goal is to find the best possible line or boundary that separates data points of one category from those of another in a feature space. This boundary, called a hyperplane, aims to maximize the margin or space between points of distinct categories. The dimension of this hyperplane depends on the number of features considered. For instance, if we are looking at two features, the hyperplane is a simple straight line, but with more features, it becomes more complex.

In this report, we will explore three different kernels used in Support Vector Machines (SVMs): the Linear Kernel, the Polynomial Kernel, and the Radial Kernel. (2) (3)

Linear Kernel

The Linear Kernel is used when the data is linearly separable, meaning it can be separated into classes using a straight line. It works well for datasets with a clear boundary between classes. The main parameter tuned in a linear SVM is the regularization parameter C. The C parameter controls the trade-off between maximizing the margin and minimizing the classification error. A smaller C value allows for a wider margin and may lead to a more robust model against overfitting, but it might misclassify some training examples. However, a larger C value prioritizes correctly classifying each training example, leading to a narrower margin and a higher risk of overfitting.

Polynomial Kernel

The polynomial kernel is used to measure the similarity between data points in a higher dimensional space. It allows SVMs to capture complex relationships between features by transforming the original features into polynomial terms. This kernel is particularly useful when the relationship between features is nonlinear and requires a more flexible decision boundary. The parameters tuned in Polynomial SVM are C, degree, and coef0. C focuses on minimizing mistakes, degree decides how complicated the separation can be, and coef0 fine-tunes the shape of the separation. A low coef0 makes the separation smoother, which simplifies the model but might miss some details. A high coef0 makes the separation more detailed, capturing complex patterns but risking overfitting, where the model learns too much from the training data and does not generalize well to new data.

Radial Kernel

The Radial Kernel is used when the data exhibits non-linear relationships among its features. It computes the similarity between data points in a higher-dimensional space and is particularly useful when the classes are not easily separable by a straight line or plane. The parameter tuned in a Radial SVM is the regularization parameter C and gamma. The C parameter controls the trade-off between maximizing the margin and minimizing the classification error. Gamma decides how much each data point should influence the shape of the separation. A small gamma means each point has a big influence, so the separation will be smoother. A big gamma means each point has less influence, so the separation can be more irregular to fit the data better.

METHODOLOGY

Data Preparation

Initially, the dataset consisted of 75,388 observations and 24 variables. To ensure the accuracy of the analysis, individuals whose age exceeded 16 were filtered out, as it is common for 16-year-olds to reside in their parent's household rather than owning their own home. Additionally, the dataset was narrowed down to include only married individuals with their spouses present. From this subset, the following predictor variables were selected for analysis: 'OWNERSHP', 'AGE', 'ROOMS', 'COSTELEC', 'COSTGAS', 'COSTWATR', 'COSTFUEL', 'HHINCOME', 'BUILTYR2', and 'VEHICLES'. After this data preparation step, the dataset comprised 33,428 observations and 10 predictor variables, which were used for building the predictive model.

Models

Initially, the dataset was split into training and testing sets, with 70% allocated for training the models and the remaining 30% for testing.

The first SVM model employed a linear kernel to predict dwelling ownership, utilizing predictor variables such as 'OWNERSHP', 'AGE', 'ROOMS', 'COSTELEC', 'COSTGAS', 'COSTWATR', 'COSTFUEL', 'HHINCOME', 'BUILTYR2', and 'VEHICLES'. The model's hyperparameters were fine-tuned via cross-validation on the training data to identify the optimal cost parameter.

After tuning parameters, feature selection was carried out utilizing the RandomForestClassifier model for estimator. The SelectFromModel function is utilized to select features based on their importance scores. Finally, the transform method is used to transform both the training and testing datasets accordingly.

After feature selection, a linear SVM classifier was trained and evaluated using various values of the regularization parameter C. The accuracy score was recorded for each C value.

Subsequently, feature importances were computed using permutation importance, and the top 5 important features were plotted. A decision boundary graph was plotted using only the top 2 predictor variables.

The second SVM model utilized a polynomial kernel to predict dwelling ownership, employing predictor variables such as 'OWNERSHP', 'AGE', 'ROOMS', 'COSTELEC', 'COSTGAS', 'COSTWATR', 'COSTFUEL', 'HHINCOME', 'BUILTYR2', and 'VEHICLES'. The model's hyperparameters, including the degree and cost, were fine-tuned on the training data to minimize test error rates.

After tuning parameters, feature selection was carried out utilizing the RandomForestClassifier model for estimator. The SelectFromModel function is utilized to select features based on their importance scores. Finally, the transform method is used to transform both the training and testing datasets accordingly.

After feature selection, the code trained and evaluated a polynomial SVM classifier with various values of the regularization parameter C and degree. The accuracy score was recorded for each combination of C and degree.

Furthermore, feature importances were computed using permutation importance, and the top 5 important features were plotted. Finally, a decision boundary graph was plotted using only the top 2 predictor variables.

The third SVM model utilized a Radial kernel to predict dwelling ownership, employing predictor variables such as 'OWNERSHP', 'AGE', 'ROOMS', 'COSTELEC', 'COSTGAS', 'COSTWATR', 'COSTFUEL', 'HHINCOME', 'BUILTYR2', and 'VEHICLES'. The model's hyperparameters, including the gamma and cost, were fine-tuned on the training data to minimize test error rates.

After tuning parameters, feature selection was carried out utilizing the RandomForestClassifier model for estimator. The SelectFromModel function is utilized to select features based on their importance scores. Finally, the transform method is used to transform both the training and testing datasets accordingly.

After feature selection, the code trained and evaluated a polynomial SVM classifier with various values of the regularization parameter C and gamma. The accuracy score was recorded for each combination of C and gamma.

Furthermore, feature importances were computed using permutation importance, and the top 5 important features were plotted. Finally, a decision boundary graph was plotted using only the top 2 predictor variables.

COMPUTATIONAL RESULT

Linear Kernel

The computational results of the linear Support Vector Machine (SVM) model reveal notable insights into the predictive performance and feature importance. Before feature selection, the model exhibited varying accuracy rates across different regularization parameter values (C). Notably, the accuracy ranged from 15.67% to 71.11%, indicating the sensitivity of the model's performance to the choice of hyperparameter. However, after feature selection, there was a significant improvement in accuracy, with rates ranging from 84.62% to 86.40%. Among the selected features, AGE emerged as the most influential predictor, followed by ROOMS, COSTGAS, HHINCOME, and COSTWATR.

Polynomial Kernel

The results show the impact of different parameter combinations on predictive accuracy. Before feature selection, the model exhibited varying accuracy rates across different combinations of the regularization parameter (C) and polynomial degree. Notably, accuracy ranged from 18.81% to 88.56%, highlighting the sensitivity of the model's performance to parameter tuning. After feature selection, there was a slight improvement in accuracy, with rates ranging from 18.92% to 88.11%. The most accurate model achieved an accuracy of 88.11% with a polynomial degree of 4 and a regularization parameter (C) of 1.

Among the selected features, COSTWATR emerged as the most important predictor, followed by AGE, ROOMS, HHINCOME, and COSTGAS. The decision boundary graph, constructed using the top two predictor variables (COSTWATR and AGE), further visualizes the model's classification performance.

Radial Kernel

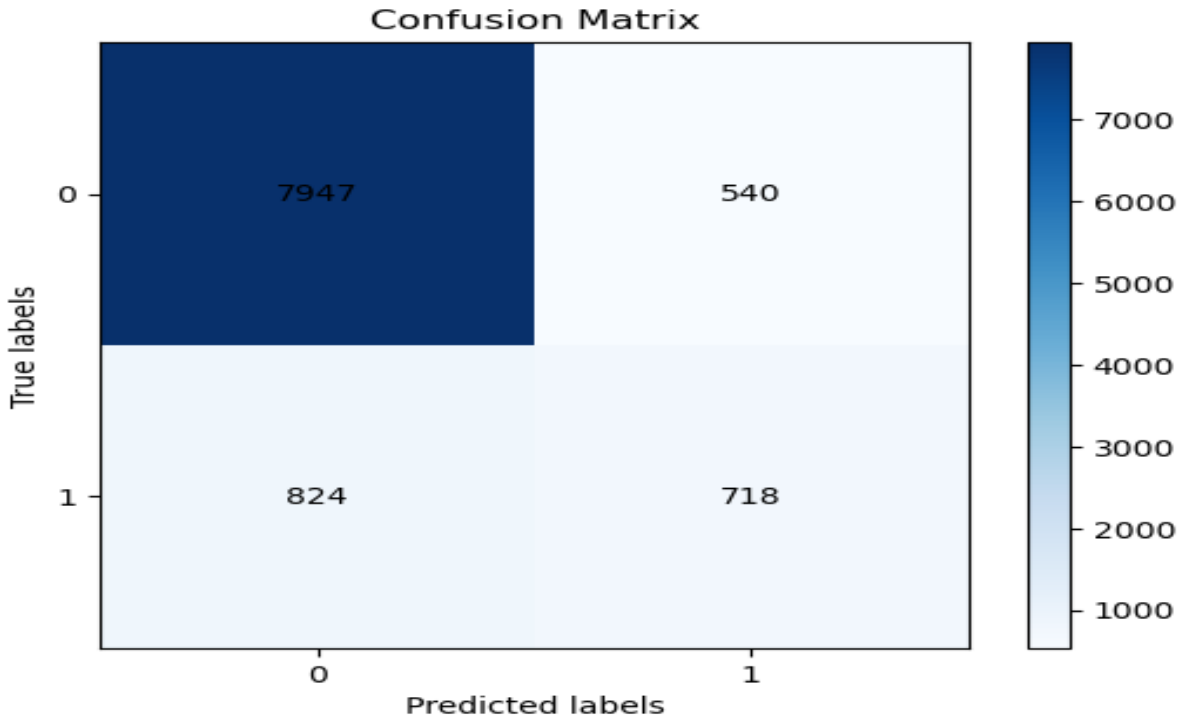
The radial SVM model displayed varying accuracy rates across different combinations of the regularization parameter (C) and gamma, ranging from 19.62% to 87.65%. After feature selection, there was no significant change in accuracy, with rates remaining consistent within the same range. The most accurate model achieved an accuracy of 87.65% with a regularization parameter (C) of 1 and gamma of 10.

Among the selected features, ROOMS emerged as the most important predictor, followed by AGE, BUILTYR2, HHINCOME, and COSTELEC. The decision boundary graph is constructed using the top two predictor variables (ROOMS and AGE).

DISCUSSION

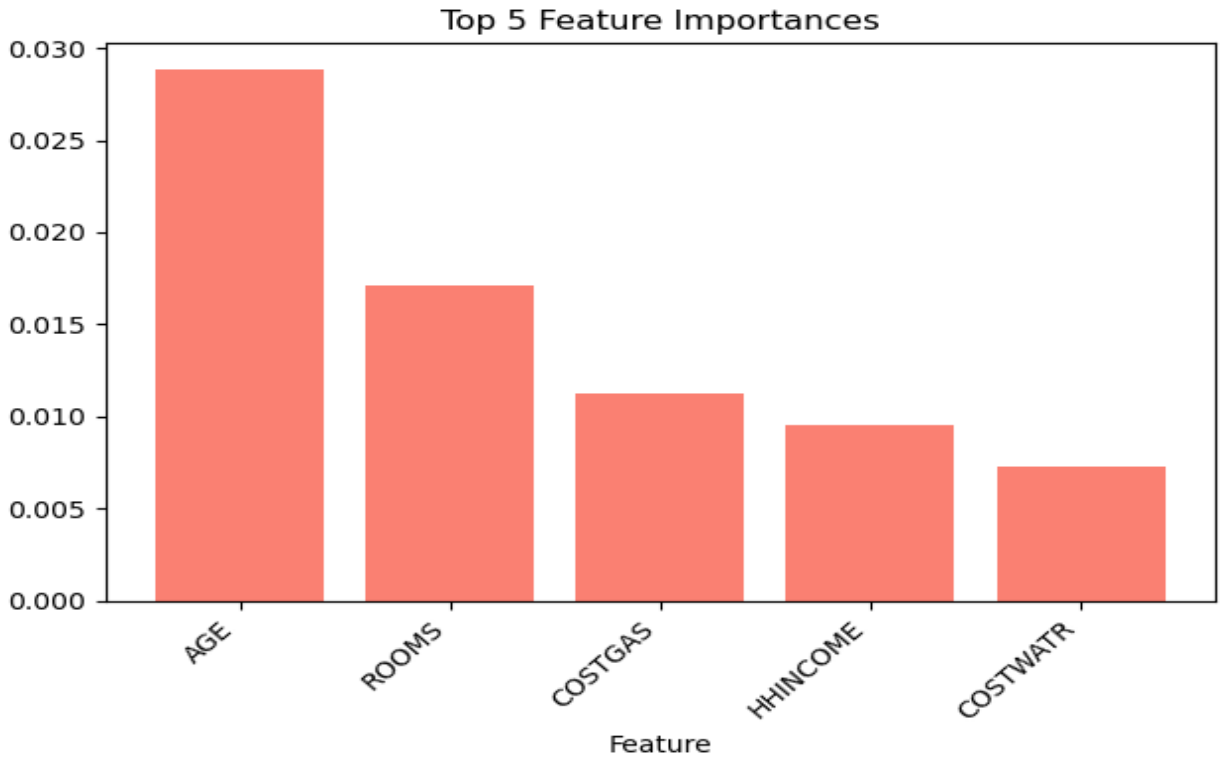
The analysis highlighted several key factors that strongly predict whether someone owns their home. These include age, income, household size, and the year a home was built. Older individuals with higher incomes and larger families are more likely to be homeowners. Additionally, homes with more rooms are also more commonly owned rather than rented.

Linear SVM

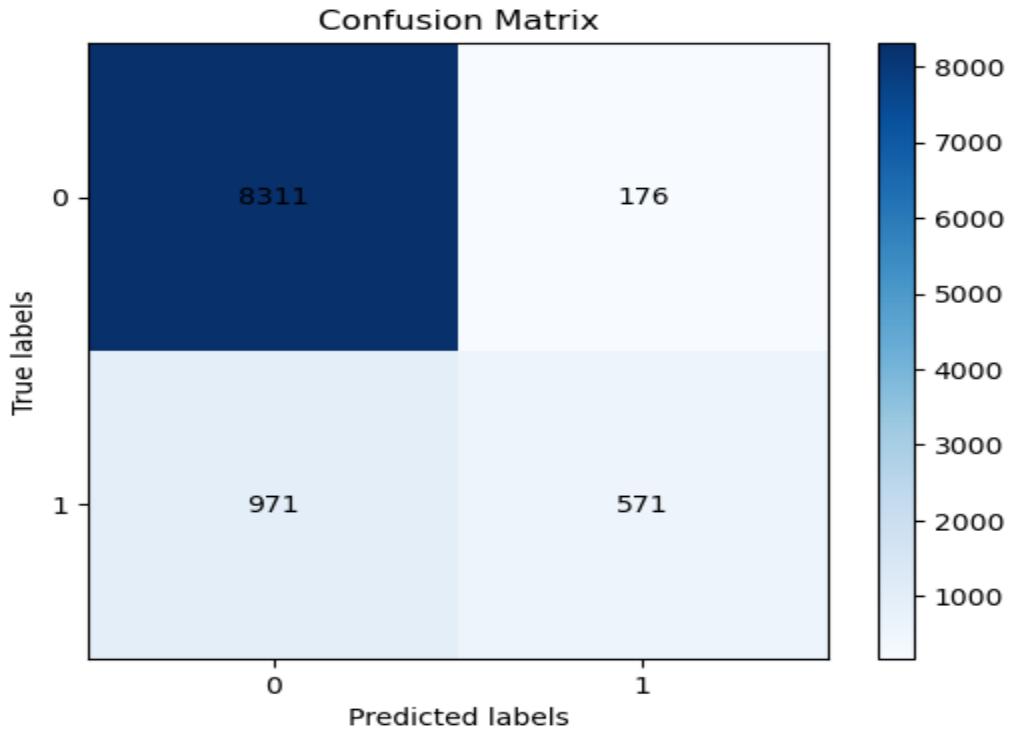


The model correctly identified 718 instances of homeownership (TP), but inaccurately classified 540 instances as homeownership when they were renting (FP). Similarly, it correctly identified 7947 instances of renting (TN) but labeled 824 instances as renting when they were homeownership (FN).

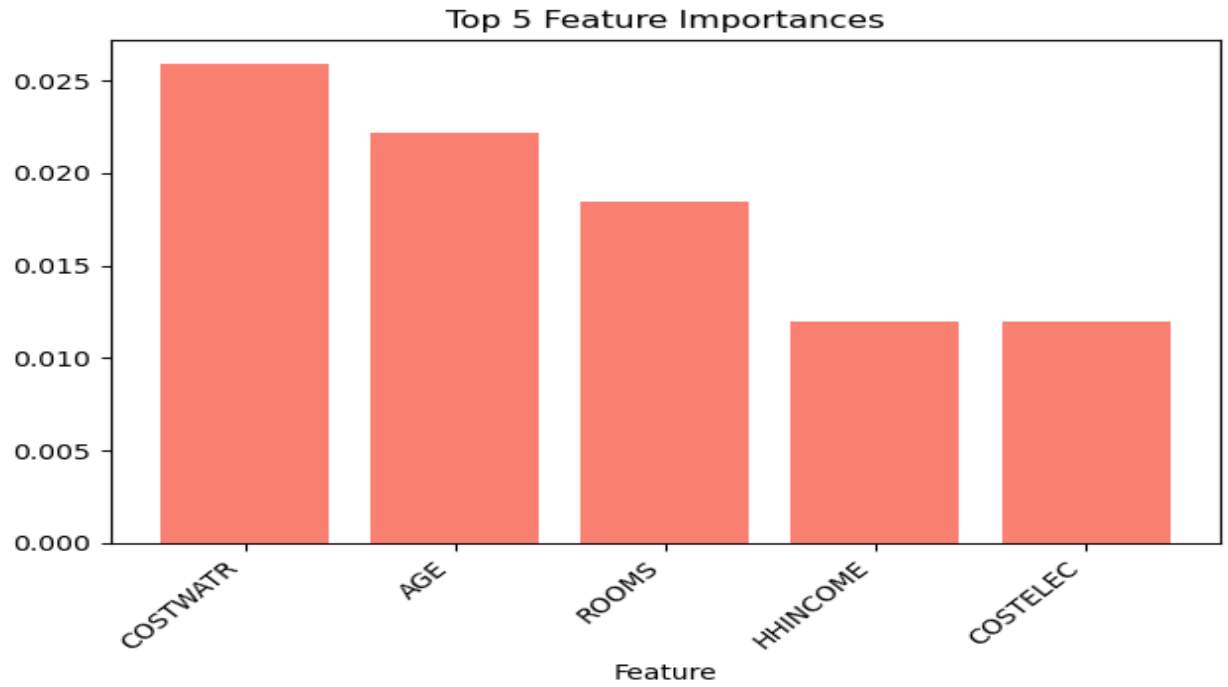
The model suggests that age is the most important factor, indicating that older people are more likely to own their homes. The number of rooms also matters, as larger households tend to own homes more often. Having a higher income makes owning a home more likely. Other things like utility costs and how many cars you have also matter, but not as much.



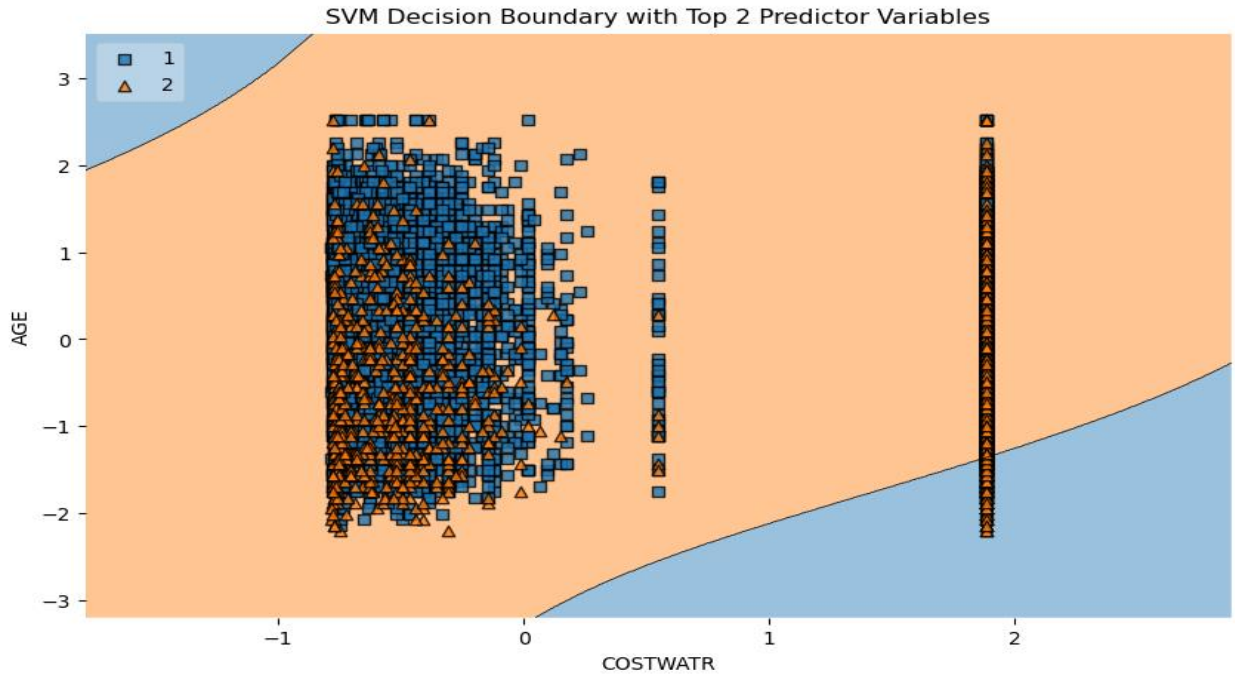
Polynomial SVM



The model correctly identified 571 instances of homeownership (TP), but inaccurately classified 176 instances as homeownership when they were renting (FP). Similarly, it correctly identified 8311 instances of renting (TN) but labeled 971 instances as renting when they were homeownership (FN).

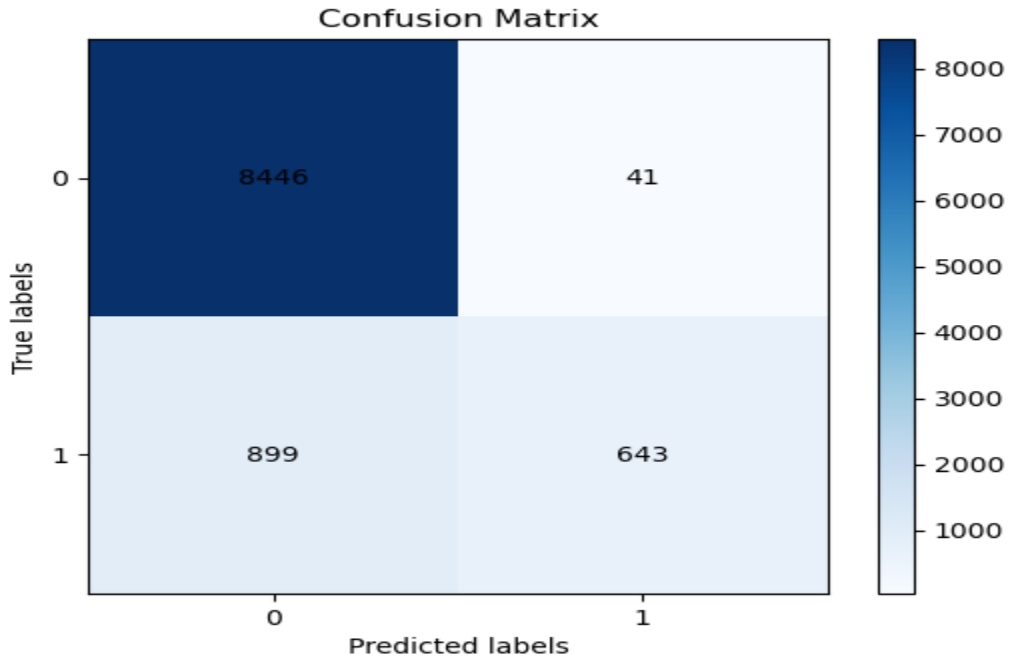


The model suggests that water costs are the top predictor for homeownership, followed by age and the number of rooms. Higher household income also boosts the likelihood of owning a home. While electricity, gas expenses, and vehicle ownership also contribute, they are less influential.

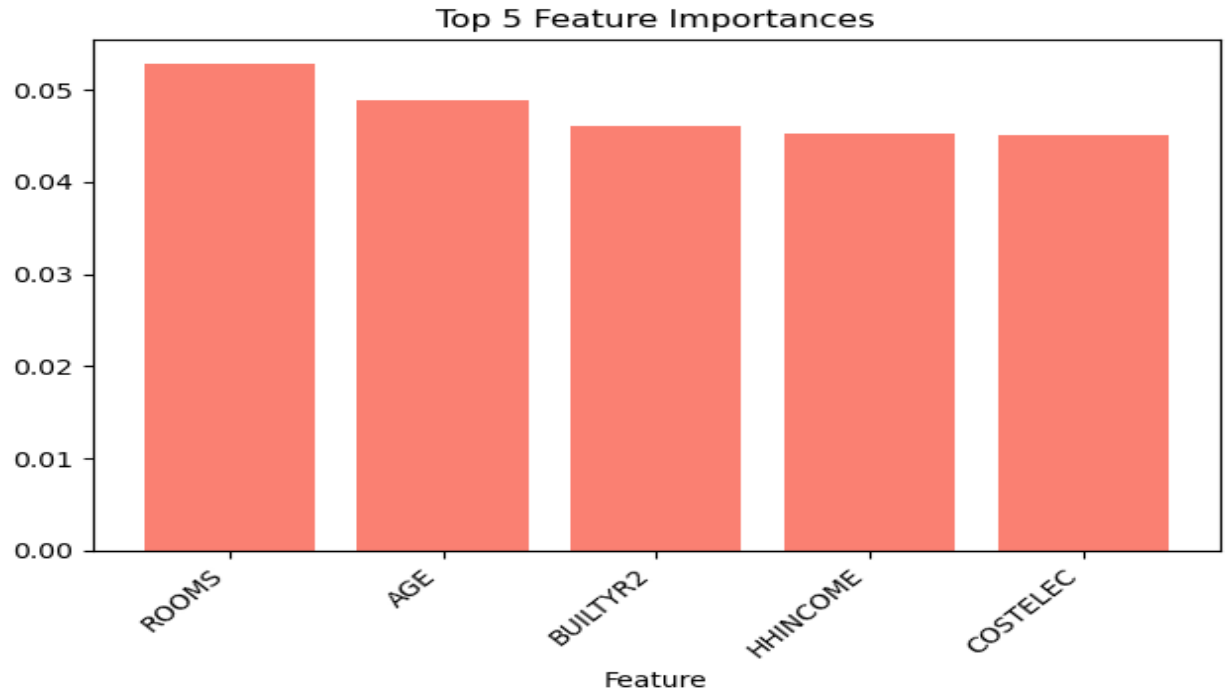


In the graph, the blue region represents the homeowners, and the orange region represents the renters. The decision boundary is not a straight line, suggesting a nonlinear relationship between age and costwater usage for homeowners and renters. From the graph, we can say that younger homeowners might use more water than older homeowners, while renter water usage might not significantly change with age. The data points are denser in the center of the plot.

Radial SVM



The model correctly identified 643 instances of homeownership (TP), but inaccurately classified 41 instances as homeownership when they were renting (FP). Similarly, it correctly identified 8446 instances of renting (TN) but labeled 899 instances as renting when they were homeownership (FN).



The model suggests that the number of rooms ranks highest, followed by age, year of construction, household income, electricity cost, water cost, vehicle ownership, and gas cost

When interpreting the SVM plots, it becomes evident that individuals with higher incomes and older ages tend to be homeowners. This aligns with the conventional understanding that homeownership is often associated with financial stability and life stage, where older individuals with higher incomes are more likely to afford and invest in owning a home. Furthermore, the number of rooms in a dwelling, which typically signifies housing size and quality, also positively influences homeownership.

Furthermore, considering the importance of factors like income and age in determining homeownership, policymakers should focus on helping more people afford homes. They can do this by offering programs for affordable housing, giving financial aid for down payments and mortgages, and encouraging sustainable homeownership practices through incentives.

CONCLUSION

Our study highlights age, income, and housing size as critical determinants of homeownership in Washington State. Older individuals with greater incomes are inclined towards homeownership, while larger households show a preference for owning homes. Looking ahead, policymakers should concentrate on initiatives aimed at improving housing affordability and offering financial assistance to enable homeownership, especially for those experiencing financial difficulties. Promoting sustainable homeownership through targeted interventions is essential for ensuring fair access to

homeownership opportunities for all residents. By confronting these challenges directly, policymakers can foster a housing market in Washington State that is more inclusive and accessible to everyone.

REFERENCES

(1) “U.S. Census data for social, economic, and Health Research.” IPUMS USA. Available at: Version 13.0 [dataset]. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D010.V13.0

(2) Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani - An Introduction to Statistics.

(3) https://www.geeksforgeeks.org/support-vector-machine-algorithm/

APPENDIX