Emma Oriol
5/1/2023

Understanding Home Ownership in Washington State

Abstract:

This paper utilizes IPUMS (Integrated Public Use Microdata Series) census data to predict renting/ownership status among married people in Washington state. Variables related to socioeconomic status and demographics of residents are used to build various support vector models that classify home ownership status. The results of these models can be used to better understand factors influencing home ownership and inform policy to improve affordable housing in Washington state.

Introduction:

Home ownership has been linked to a range of social and economic benefits to owners, including increased wealth, improved health, and higher education achievement (Yun & Evangelou, 2016). The goal of this report is to understand what factors are most important in predicting home ownership in order to inform policy decisions that increase access to housing in Washington.

This report uses IPUMS census data to develop predictive models for home ownership/rental status in Washington state. The data contains information on more than 18,000 distinct households composed of married individuals and contains variables relating to socioeconomic status and demographics, including household income, age, education level, utilities cost, population density, and house size. Various subsets of these variables are used to build three support vector models (one linear, one radial, and one polynomial) that predict home ownership status.

Theoretical Background:

Support vector models (SVMs) classify data by computing a hyperplane that best separates two classes of data points. The support plane has a margin on either side that helps maximize the distance between the hyperplane and classes. Any points that lie within the margin or are misclassified by the hyperplane are considered “support vectors”. Only support vectors influence the fit of the hyperplane. SVMs have several advantages, including being efficient for large data sets because only data points that are support vectors are needed to generate the model. Another advantage is that SVMs have some flexibility in that they allow points to be misclassified. One disadvantage of SVMs is that they can be oversensitive to the training data, especially if there are too few support vectors (James et al., 2022).

Three types of support vector models are used in this report; linear, polynomial, and radial. Linear SVMs compute a linear hyperplane to classify data. Polynomial models compute a curved plane by raising the linear relationship to the nth degree, specified by the parameter ‘degree’. Lastly, radial models allow even more curved and circular boundaries. They use a parameter called ‘gamma’ to increase the Euclidean distance between points and give more weight to close together points. As gamma increases, the sphere of influence around each point decreases (James et al., 2022).

One parameter of linear, polynomial, and radial SVMs is ‘cost’, which determines the number of data points allowed within the margin. A high value of cost means that there is a high allowance for data points to be within the margin. Therefore, as cost increases, the margin will increase and include more support vectors. The model is tuned to find the optimal value of cost by trying several different values of cost and choosing the one with the lowest training error (James et al., 2022).

For polynomial models, the optimal degree is tuned in conjunction with cost. A list of degree values are passed to the model with cost values, and the cost and degree combination that has the lowest error is selected as the optimal model. Similarly, radial models tune gamma and cost in the same way to determine the optimal radial model parameters (James et al., 2022).

Methodology:

A subset of 23 variables relating to socioeconomic status and demographics of Washington residents were chosen to be evaluated. These variables were subset in order to simplify the data and include only the most relevant variables of interest. The data was further filtered to include only the oldest member in each household in order to examine only unique households represented by an assumed ‘head of household’. Lastly, the data was filtered to only include married people in order to further reduce the dataset, which was initially too large to process. Married people were chosen as a relevant group to evaluate, because home ownership is most common among married people (Yun & Evangelou, 2016).

A random sample of 180 observations of the data (without replacement) were chosen each for training and testing sets. This sample size was used because it was the largest sample size that could be computed without crashing the program.

The first support vector model is a linear model that predicts home ownership status based on the selected predictor variables of age of oldest household member and total household income. Tuning was used to select the optimal cost value of 1e-05. The predictive power of the model was then evaluated using the training and testing data.

The second support vector model is a radial model. In order to improve the predictive performance of the model as compared to the linear model, all variables were used as predictors. The data was first subset to remove variables relating to home ownership, marriage status, and serial number because they were not meaningful for prediction. The model was then tuned to find the optimal values for cost and gamma, which were determined to be a cost of 1000 and a gamma of 1e-05. The accuracy of the model was evaluated by calculating the training and testing classification error rate.

Lastly, the third model is a polynomial model that uses the number of vehicles owned, household income, population density, and age of oldest household member as predictors. The values for cost and degree were determined through tuning to be 10 and 1, respectively. The model was then evaluated on the training and test data sets, and the classification error was determined.

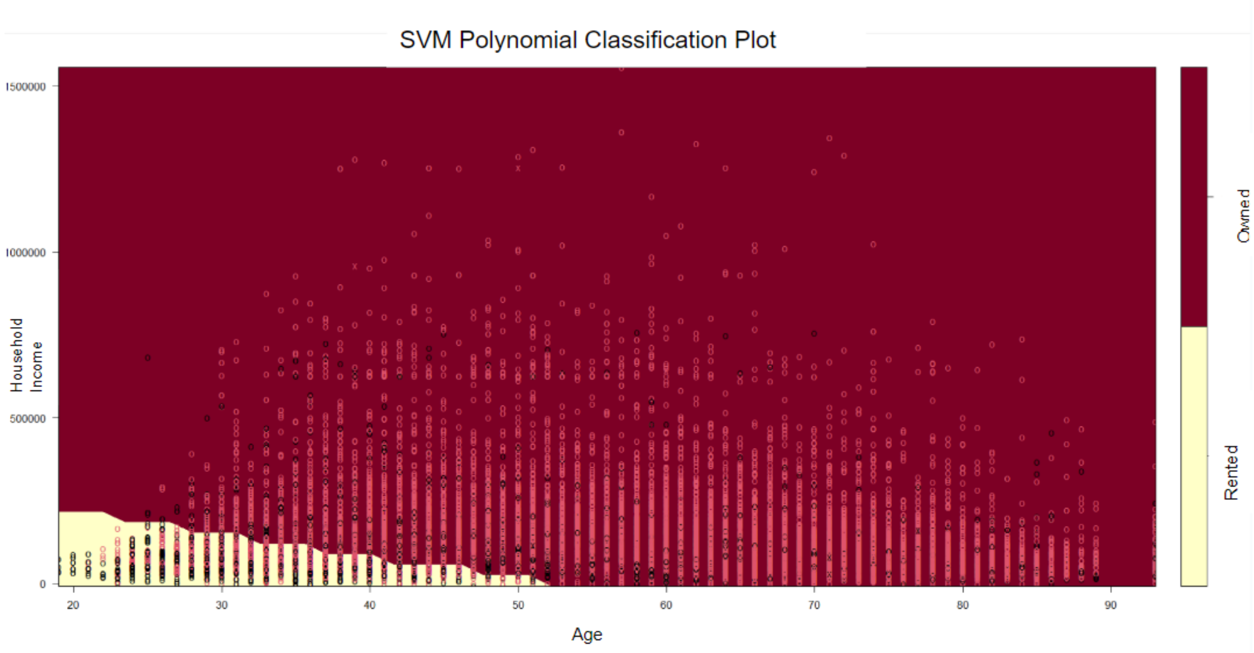
Computational Results:

The linear support vector model predicted home ownership status on the training data with an error rate of 13.5% and on the test data with an error of 15.0%. However, the model only predicted that the home was owned and never predicted rented for both the training and testing data. This could indicate that the predictor variables selected (age of oldest household member and household income) are not good predictors of home ownership or that the linear model was not an appropriate model for the data. Either way, this model is not very useful because it only predicts one category.

The radial support vector model predicted home ownership status on the training data with an error rate of 11.7% and on the test data with an error rate of 13.5% using all variables as predictors. This model showed improved prediction accuracy as compared to the linear model and predicted both categories “rented” and “owned”. When examining the coefficients of each predictor variable, the variables with the highest positive values were number of bedrooms, number of rooms, and age. This means that as these variables increase, the likelihood of home ownership increases as well. The variables with the largest negative values were number of families living in the household, cost of water, and cost of fuel. This indicates that as these variables increase, the likelihood of ownership decreases.

Lastly, the polynomial support vector model predicted home ownership status on both the training data and test data with an error rate of 12.8% using the predictor values of age, household income, vehicles, and population density. This model also predicted both ‘rented’ and ‘owned’ values. The coefficients of each predictor show that age and household income were the strongest predictors of home ownership, with the likelihood of ownership increasing as both age and household income increase. The number of vehicles increasing owned also increased the likelihood of ownership, but to a lesser degree, and the population density increasing slightly decreased the likelihood of ownership.

The classification plot for this polynomial model is shown below. The red and beige colors represent the decision boundary for rented vs owned, respectively. The bottom left corner, where age and income are the lowest, shows where the model will predict ‘Rented’. For all other data points that fall in the red area, the model will predict ‘Owned’.



Discussion:

The models showed several strong predictors associated with higher likelihood of home ownership. Larger homes (i.e. homes with higher numbers of rooms and bedrooms) were more likely to be owned, and older people with high household income were more likely to own homes. This demonstrates that there is a lack of affordable housing in Washington by showing that homeownership is much more likely for high income people looking for larger homes. Therefore, there are not enough smaller, affordable homes for low income people and young people.

The models also showed several predictors are associated with decreasing likelihood of home ownership, such as cost of gas and water. This suggests that high utility costs are a barrier to people owning homes. Additionally, homes with multiple families living in them are less likely to be owned. One explanation is that households have multiple families living in them due to people with limited financial resources living together to share the burden of living costs. This again demonstrates that low income people are much less likely to own homes.

In conclusion, these findings highlighted that low-income and young people in Washington have low home ownership rates, and that high utility costs and lack of affordable housing are barriers to home ownership. Washington policy makers should consider implementing government programs to increase the availability of smaller and more affordable homes and lower utility costs, especially for young and low-income people. Example policies could include income-qualified tax breaks for home ownership, low-interest mortgages for first-time home owners, and government-subsidized utility programs. These programs would promote equity by making home ownership more accessible to all Washingtonians, despite age and income.

Bibliography:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2022). *An introduction to statistical learning: With applications in R*. Springer.

Minnesota Population Center. (n.d.). U.S. Census data for social, economic, and Health Research. IPUMS USA. Retrieved April 30, 2023, from https://usa.ipums.org/usa/

Yun, L., & Evangelou, N. (2016, December). *Social benefits of homeownership and stable housing - gmar*. Greater Milwaukee Association of Realtors. Retrieved April 26, 2023, from https://www.gmar.com/data/resources_files/Social%20Benefits%20of%20Homeownership%20%20Stable%20Housing.pdf