**Predicting Dwelling Occupancy based on Age, Marital Status and Utilities**

**Abstract:**

This study employs Support Vector Machine techniques to predict whether a dwelling is occupied by the owner or renter based on demographic and housing variables. Using data from Washington State collected from the US Census, variables such as age, marital status, vehicles owned, and utility costs are analyzed. SVM models with Linear, RBF, and Polynomial kernels are utilized for classification. Performing on the different kernels and cross validating models drew 84 percent of accuracy in categorizing the data. The findings provide insights for real estate agents and policymakers into housing occupancy patterns and the influence of demographic and housing factors on predictions.

**Introduction:**

This research is used to predict the dwelling occupancy in order to know whether it is owned or rented property based on the factors such as age, vehicles, marital status and utilities bills. The data for this research is collected from Washington State data collected from US Census has more than 70000 columns. For analyzing the data Support Vector Machine with different methods like Linear, Radial Basis Function and Polynomial Kernel. In this report all the three models are discussed and the insights from the analysis are helpful for real estate agents and policymakers to understand how factors influenced the dwelling occupancy.

**Theoretical background:**

Support Vector Machines (SVMs) is applied to regression and classification problems. When there are more features than samples and high-dimensional spaces, they are especially useful. SVMs seek to maximize the margin between classes while identifying the ideal hyperplane in the feature space that divides them. We use SVMs for binary or multi-class classification problems, particularly where there is a need for a distinct decision boundary or when dealing with high-dimensional data.

Decision Boundary: A straight line dividing the classes serves as the decision boundary when dealing with linearly separable data. The data points that are closest to the decision boundary ,the support vector machines, determine this line.

Margin: The separation between the support vectors and the decision boundary. To improve generalization to data that has not yet been seen, the maximal margin classifier chooses the hyperplane with the biggest margin.

Kernel:A kernel in Support Vector Machines is a function that calculates the inner product or similarity measure between pairs of data points, simplifying non-linear classification

Types of Kernels:

For Linear Kernel: The inner product between the input features. Used for linearly separable data or when computational efficiency is essential. To visualize data points and decision boundary as a straight line Effective computation, particularly with big datasets. Fits data that can be separated linearly.Restricted to decision boundaries that are linear. Less useful for data that are not linearly separable.The linear kernel is easy to use because it doesn't require tuning of any additional parameters.

For Radial Kernel:Using a Gaussian distribution, the radial kernel also referred to as the Gaussian kernel transforms data into a higher dimensional space.Good for documenting intricate decision limits and non-linear interactions. The flexibility to represent intricate, non-linear patterns is one of its.Plots decision boundary as a smooth curve, visualizing the effect of different values of gamma on the decision boundary.Beneficial for data with several dimensions. Requires more computing power than a linear kernel, particularly when dealing with big datasets. Dependent on the kernel width parameter (gamma) selection.The most important parameter, gamma , establishes the impact of every training sample; larger values yield more intricate decision limits.

For Polynomial Kernel:This kernel transforms data into higher dimensional space using polynomial functions.Suitable for capturing non linear relationships in a polynomial form.Decision boundary as curved surface shows the impact of the polynomial degree on the complexity of the boundary.Captures complex,non linear decision boundaries. It allows fine tuning of the degree parameter to control model complexity.More expensive computationally and sensitive .The degree of the polynomial. It determines the complexity of the decision boundary.

Metrics for Success:

Classification Accuracy: Calculates the percentage of cases that were successfully classified.

Confusion Matrix: Assess the quantity of true positives, true negatives, false positives, and false negatives.

**Methodology:**

The aim of the study is to predict the dwelling occupancy whether it is by owner or renters. In order to predict the data is taken in from the US census in Washington state. This data set has multiple columns which are uncleaned and coded. To make the data useful for the analysis the data needed to be processed.

Data Processing: The data is grouped by the serial number to have an order in the data to proceed with further processing. Once the data is grouped, irrelevant columns being dropped. In that detailed information of ownership as it is corelated to the other column ownership of the dwelling ,number of families living in, the education column is dropped ,the year the building was built in , the high income column is also dropped as it is corelated to the total income column .The unique columns is printed for better understanding. Now the total income is set to mean and the age is choose to maximum. Now as few columns are coded , those columns are modified with the help of code book. The first dealt in modifying is vehicles column using lambda function if the value is equal to it is replaced with 0 otherwise it is kept to the original value. The first column represents individuals who are married ,values 1 or 2 in marital status, the second column represents individuals who are divorced ,values 3, 4, or 5 in marital status, and the third column represents individuals who are not married ,value 6 in marital status. For the utilities columns which includes electricity, water ,gas and fuel , for these columns if the value is greater than 9992 it leads to outliers, so it is replaced with 0 otherwise it is kept to the original value. These new encoded values are stored into new columns . And after this processing the marital status and utilities columns are dropped as the new columns are already created.

Splitting dataset: This preprocessed data set is divided into predictor variables and the target variables. In predictor variables age, bedrooms in the residence, number of vehicles, the newly creates marital status and newly encoded utilities columns. The house ownership is set to be target variable. Using Min Max scaler the processed data is scaled after splitting training and testing dataset.

Support Vector Machine modelling:

The support vector machine modelling with linear, radial and polynomial kernel. For all the kernels following steps are followed.

1. Parameter Tuning: Grid search is employed to tune the regularization parameter 'C' for the linear SVM model. A predefined grid of values for 'C' is specified, ranging from 0.001 to 100. The GridSearchCV function from scikit-learn is utilized with 5-fold cross-validation to find the best 'C' value that minimizes the cross-validation error rate.

2. Model Fitting: After parameter tuning, a new linear SVM model is instantiated with the best 'C' value obtained from the grid search. This model is then fitted to the training data (X\_train) and corresponding labels (y\_train).

3. Decision Boundary Visualization: To visualize the decision boundary of the SVM model, a mesh grid is created spanning the range of the two most important features in the training data. Predictions are made for each point in the mesh grid, and a contour plot is generated to display the decision boundary separating different classes.

4. Model Evaluation: The trained SVM model is evaluated on the testing data (X\_test) to assess its performance. Predictions are made using the tuned model, and a confusion matrix is computed to summarize the model's classification results. Additionally, the test error rate and test accuracy are calculated to quantify the model's predictive performance.

By following this methodology, the code systematically tunes the SVM model's hyperparameters, fits the model to the training data, visualizes the decision boundary, and evaluates the model's performance.

**Computational Results:**

For linear kernel:

The value 100 was identified through GridSearchCV, which explored a range of C values defined in `grid\_params` and selected the one yielding the lowest cross-validation error rate. The 0.18% cross validation error rate indicates strong model performance on unseen data based on the cross-validation splits employed during hyperparameter tuning.

Test set performance: The test error 17.61 % this represents the percentage of samples the model misclassified on the unseen test set (X\_test, y\_test). The test accuracy 82.39% this value is calculated as 1 - test error and signifies the percentage of samples correctly classified on the test set.

Confusion Matrix: The confusion matrix is:



Here Class 1 represents owners and Class 2 represents renters

True Positives: 9067 instances were correctly classified as belonging to Class 1.

False Negatives: 434 samples from Class 1 were incorrectly identified as Class 2 (missed detections of Class 1).

False Positives: 1854 samples from Class 2 were mistakenly classified as Class 1 (misidentified Class 1).

True Negatives :1641 instances were correctly classified as belonging to Class 2.

Visualization: A decision boundary plot was generated to visualize how the trained model separates the two classes in the feature space. This plot focused on the two most significant features age and the number of bedrooms in the residence.

A diagram of a different color scheme

Description automatically generated with medium confidence

Fig (a) Decision Boundary for Linear kernel

The decision boundary separates the data points into two regions. The data points on one side of the boundary are classified as one class, and the data points on the other side of the boundary are classified as another class.The decision boundary does not perfectly separate the two classes. There are some data points from one class that lie on the wrong side of the decision boundary.

For RBF kernel:

Grid Search for Optimal Parameters:The code employs GridSearchCV to find the best combination of hyperparameters for the Support Vector Classifier (SVC) with an RBF kernel. It evaluates various settings for the regularization parameter (C) and the gamma parameter, which controls the influence of data points in the kernel function.

Best Parameters Identified:The grid search identified C=1 and gamma=5 as the hyperparameter combination yielding the lowest cross-validation error rate. This implies that an SVC model with these parameters performed well on the data while splitting it into training and validation sets during the grid search process.

Cross-Validation Error Rate: With the chosen hyperparameters (C=1, gamma=5), the model achieved a cross-validation error rate of 0.17%. This indicates strong generalizability of the model on data based on the cross-validation splits. The test accuracy is of 83.5% .

Confusion Matrix: The confusion matrix is:



Here Class 1 represents owners and Class 2 represents renters

True Positives: 9012 instances were correctly classified as belonging to Class 1.

False Negatives: 489 samples from Class 1 were incorrectly identified as Class 2 (missed detections of Class 1).

False Positives: 1644 samples from Class 2 were mistakenly classified as Class 1 (misidentified Class 1).

True Negatives : 1851 instances were correctly classified as belonging to Class 2.

Visualization: A decision boundary plot was generated to visualize how the trained model separates the two classes in the feature space. This plot focused on the two most significant features age and the number of bedrooms in the residence. The decision boundary is not a straight line, but rather a curved line. This suggests that a more complex relationship exists between the number of bedrooms, age of the house, and dwelling occupancy.

A diagram of a bed

Description automatically generated

Fig (b) Decision Boundary for RBF kernel

Features: The plot focuses on two features: number of bedrooms and age . These were likely chosen because they might be informative for predicting dwelling occupancy by owners or renters.

Decision Boundary: The curved line in the plot is the decision boundary learned by the model. It separates the data points into two regions.

Data Points and Colors: The data points are colored based on the actual class labels. This allows us to visualize how well the decision boundary separates the two classes.

For Polynomial kernel:

Grid Search for Optimal Parameters:The code employs GridSearchCV to find the best combination of hyperparameters for the Support Vector Classifier (SVC) with a Polynomial kernel. The grid search evaluated different values for the regularization parameter (C) and the degree of the polynomial kernel.

Best Parameters Identified: The grid search identified C=100 and degree=5 as the hyperparameter combination yielding the lowest cross-validation error rate. This implies that an SVM model with these parameters performed well on the data while splitting it into training and validation sets during the grid search process.

Cross-Validation Error Rate: With the chosen hyperparameters (C=100, degree=5), the model achieved a cross-validation error rate of 0.17%. This indicates strong generalizability of the model on unseen data based on the cross-validation splits. The test accuracy is of 84%

Confusion Matrix: The confusion matrix is:



Here Class 1 represents owners and Class 2 represents renters

True Positives: 9096 instances were correctly classified as belonging to Class 1.

False Negatives: 405 samples from Class 1 were incorrectly identified as Class 2 (missed detections of Class 1).

False Positives: 1723 samples from Class 2 were mistakenly classified as Class 1 (misidentified Class 1).

True Negatives : 1772 instances were correctly classified as belonging to Class 2.

Visualization: The curved line in the plot is the decision boundary learned by the model. It separates the data points into two regions. Houses on one side of the decision boundary are classified as one class , and houses on the other side are classified as the other class .

A diagram of a different color scheme

Description automatically generated with medium confidence

Fig (c) Decision Boundary for Polynomial kernel

The decision boundary is not a straight line, but rather a complex curve. This suggests that a non-linear relationship exists between the number of bedrooms, age, and dwelling occupancy. A linear SVM model might not have been able to capture this complexity as effectively as the SVM with a polynomial kernel. There seems to be some overlap between the two classes near the decision boundary.

**Discussion:**

The study aimed at intricate relationship between various socioeconomic factors and homeownership, focusing on age, income, utilities and marital status as primary determinants. In order to predict whether the residence is occupied by the owner or the renter. For predicting Support Vector machines are used by using 3 types of kernels. For all kernels we had target variable as ownership and predicting variables as age, bedrooms , vehicles, marital status and utilities bill. In linear kernel we got lowest test accuracy compared to RBF and polynomial kernels. There was not such major difference in test accuracy between radial and polynomial but the polynomial had high test accuracy with 84 % which shows it had high model performance.

Additionally, the study struggled with several intrinsic constraints of the dataset and analytical methodology. The difficulty of establishing temporal correlations among variables was highlighted, which hindered causal inference. Due to the high correlation between the variables, the dependence on self-reported data also created potential biases and restrictions, which contributed to data loss. Not withstanding these limitations, the analysis offers insightful information about the socioeconomic factors that influence home ownership, helpful for further studies that will address methodological issues and examine more complex correlations. Overall, even if the study's findings add significantly to understanding of housing dynamics, further work is needed to improve analytical techniques and increase the number of variables in the dataset so that the results are more reliable and applicable to larger populations.

**Conclusion:**

In summary, research has examined the complex interplay between socioeconomic factors and house occupancy, highlighting the critical roles played by age, income, utilities, and marital status. Using a wide range of housing and demographic data, attempted to determine whether a dwelling is owned or rented through Support Vector Machine (SVM) techniques. Investigation has provided detailed information about the predicted performance of SVM models, with the Polynomial kernel appearing as the most effective model with an 84% test accuracy. These findings have significant ramifications and provide stakeholders with a better grasp of the dynamics of home occupancy and the complex effects of socioeconomic determinants.

Observing previously, it is clear that the work has not only shed light on the factors that influence home ownership, but it has also demonstrated how sophisticated modeling methods can be used to analyze intricate socioeconomic phenomena. The plan to keep improving and exploring in the future in an effort to broaden analytical perspectives and add more to the conversation on housing dynamics. The ultimate goal is to equip real estate players and policymakers with the knowledge necessary to properly traverse the always changing housing dynamics landscape and make well-informed decisions.

**References:**

[1] Steven Ruggles, Sarah Flood, Matthew Sobek, Danika Brockman, Grace Cooper, Stephanie Richards, and Megan Schouweiler. IPUMS USA: Version 13.0 [dataset]. Minneapolis, MN: IPUMS, 2023. <https://doi.org/10.18128/D010.V13.0>

[2] Creating Linear kernel plots- <https://www.geeksforgeeks.org/creating-linear-kernel-svm-in-python/>