Tenant Harassment Prediction: Comparative Analysis of Different Machine Learning Classifiers

Otasowie Owolafe, Samuel A. Adegbite
Federal University of Technology, Nigeria

Abstract

Occupant harassment is a typical occurrence influencing many individuals all over the globe. Notwithstanding, a lot less privileged individuals have had un-told difficulties from covetous and gain-loving landowners or agents who utilize different strategies to make life uncomfortable for them as occupants. This research, therefore, carries out a comparative analysis of developed machine learning algorithms to foresee the chance of an occupant being harassed and the learning model that performs best. Online dataset about tenant's provocation having eleven attributes was downloaded. Relevant features were extracted using the lasso algorithm while Naive Bayes, Logistic Regression and Support Vector Machine (SVM) was utilized for anticipating whether or not there is harassment. The outcome shows that Logistic Regression performs better than the other classifiers with a precision of 95.4% while Naive Bayes has an accuracy of 94.7%, with SVM having minimal exactness of 62%.

Keywords: Tenant, Harassment, Naïve Bayes, Rented Property, Logistic Regression, Support Vector Machine

1. Introduction

A tenant is a person who uses the land or property rented from the landlord. According to the legal aid society NYC, harassment means any act or negligence by a landlord or landlord agent which results or intends to cause anyone who has the right to stay in a room or apartment to surrender his room or apartment or title to his residence. Tenant harassment is the voluntary creation of conditions by the landlord or his representatives to make the tenant to be guilty of breaching the rent agreement or in any case leave the hired property currently occupied by the tenant [1].

Harassment against tenants includes removing services (e.g., removing heater), intimidating tenants with ejection notice, refusing to make essential repairs, or trying to prevent the tenants from having access to the facility [2].

Tenant harassment is a common occurrence affecting many people around the world. Presently in Nigeria there is increase in demand for accommodation which is basically due to a rising population, despite various government programs and interventions, the housing shortage is estimated at 12 to 16 million [3], requiring mortgage financing that is conservatively worth between 15 and 20 trillion naira. In contrast to the homeownership situation in the US, namely 72%, 78% in the UK, 60% in China, 54% in Korea, 92% in Singapore, only about 25% of Nigerians own a house [4], making the remaining 75% available for the rental market. Many tenants are often faced with the problem of not measuring up with the expectations of the landlords. In other to guarantee the sustenance of affordable housing, several states in Nigeria have implemented some guidelines, such as rent stabilization, which limits annual rent increases, and voucher programs, in order to support low-income households. Some researchers have introduced machine learning models to help predict tenant who faces harassment from their landlord. This machine learning model is known as the predictive model. The model carries out the forecasting of previously un-classified datasets using the model that have been trained with known datasets [5]. Common predictive models are Naïve Bayes, decision trees, logistic regression, and so on.

The aim of this research is to develop a predictive algorithm like Decision tree, Logistic Regression, Naïve Bayes, and Bayesian model for predicting tenant harassment, and an algorithm to analyze principal components to reduce attributes.

2. Related Work

Several classifiers such as Logistic Regression, Random Forest, Decision Trees have been tested based on their accuracies to identify individuals who may be at higher risk of harassment.

Rebecca [6] developed a prediction system for identifying ten-ants at risk of landlord harassment using four data mining techniques as Decision Trees, Random Forest, Logistic Regression, and Gradient Boosting. The result obtained after prediction using the TSU (Tenant Support Unit) outreach data was compared with TSU internal process for finding cases and a majority of the models outperformed the TSU process. Random Forest and Gradient Boosting were the best performing models. Gürcan [7] studied the problem of text classification. It is suggested that

Copyright © 2022, Infonomics Society | DOI: 10.20533/ijicr.2042.4655.2022.0138

natural language processing methods can be helpful in text document categorization. He analyzed Turkish text classification based on supervised learning models with the parameter's variations. He considered the classification based on economy, politics, sport, health, and technology. The algorithms used are multinomial naïve Bayes, Bernoulli naïve Bayes, SVM, KNN and decision tree. The Naïve Bayes probability model was found to be more effective.

Kumar et al. [8] studied the factors affecting stock prediction using machine learning techniques. They have developed five models. These are based on SVM, random forest, KNN, naïve Bayes and Softmax. Their results indicate that the random forest algorithm and naïve Bayes are better in classification for large and small datasets respectively.

Riyanto Jayadi [9] developed an employee performance prediction using Naïve Bayes. The result shows that Naïve Bayes successfully correctly classified instances as high as 95.48%.

Palaniappan et al. [10] developed heart disease prediction system using three data mining techniques. The research used the Cleveland Heart disease Database for the prediction, the result showed that Naïve Bayes performed better than the other classifier. Also, from the result it was gathered that the relationship obtained between the attributes when Neural Network is used is more difficult to understand than that of the other models.

Ye et al. [11] used machine learning to help vulnerable tenants in New York City with Random Forest (RF), Logistic Regression (LR), Decision Trees (DT), and Gradient Boosting (GB). The evaluated result showed that Gradient Boosting performed better than Random Forest, Logistic Regression, and Decision Trees.

Molina-Gil [12] developed a system to detect harassment using machine learning and fuzzy logic techniques. The fuzzy logic model was used to detect cyberbullying signs using some linguistic input variables. The Machine learning system was used to predict the possibility of a user being victim of cyberbullying.

Masoom et al. [13] developed a system to identify bullying using four machine learning algorithms. They made use of dataset collected from college students through questionnaires to identify bullies so the authorities and the concerned guardians can proffer solutions. After testing each of the algorithms with the dataset, the result showed that Logistic Regression with an accuracy of 95.1% performed better than K-Nearest-Neighbor, SVM, and Random Forest Regression.

Yu et al. [14] developed a crime forecasting system using data mining techniques. Their first approach to architecting datasets from original crime records. The datasets contain aggregated counts of crime and crime-related events categorized by the police department. The location and time of these events is embedded in the data. Additional spatial and temporal features are harvested from the raw data set. Second, an ensemble of data mining classification techniques is employed to perform the crime forecasting. They analyzed a variety of classification methods to determine which is best for predicting crime "hotspots".

Md. Milon et al. [15] developed a system to predict breast cancer using support vector machine and k-nearest neighbors. The breast cancer termed as Wisconsin breast cancer diagnosis data set is taken from UCI machine learning repository. The performance of the proposed system is appraised considering accuracy, sensitivity, specificity, false discovery rate, false omission rate and Matthews correlation coefficient. The approach provides better result both for training and testing. Furthermore, the techniques achieved the accuracy of 98.57% and 97.14% by Support Vector Machine and K-Nearest Neighbors individually along with the specificity of 95.65% and 92.31% in testing phase.

Li et al. [16] developed a system to predict motor vehicle crashes using Support Vector Machine models. The objective of this study was to evaluate the application of SVM models for predicting motor vehicle crashes. To accomplish the objective of this study, NB regression and SVM models were developed and compared using data collected on rural frontage roads in Texas. The study showed that SVM models predict crash data more effectively and accurately than traditional NB models.

Anuja et al. [17] developed a system to classify diabetes disease using Support Vector Machine. They used datasets for diabetes disease from the machine learning laboratory at University of California, Irvine. The proposed method uses Support Vector Machine (SVM), a machine learning method as the classifier for diagnosis of diabetes. The machine learning method focuses on classifying diabetes disease from high dimensional medical dataset. The experimental results obtained show that support vector machines can be successfully used for diagnosing diabetes disease.

Axelberg et al. [18] used support vector machine for classification of voltage disturbances. They proposed a novel method based on the SVM algorithm for classification of common types of voltage disturbances. The results from the conducted experiments have shown high classification accuracy, implying that the SVM classification technique is an attractive choice for classification of this type of data.

3. Methodology

To solve tenant harassment problem, Naïve Bayes model, and Logistic Regression model would be used to predict tenant harassment by evaluating the probability based on some datasets of attributes, prediction of the probability, and evaluate the prediction accuracy.

3.1. Datasets Collection

The dataset used in this research was gotten from data. World and stored as a CSV file. The dataset consists of consecutive tenant complaints with harassment risk factors and attribute. The dataset has eleven (11) attributes amongst sex, type of complaints and the likes. Pre-processing was carried out to remove redundant data. The dataset was divided into 70% for training and 30% for testing.

3.2. Feature Selection

This is the process of selecting a subclass of features that are vital for the development of the model, it involves the removal of features that are not essential in the development of the model. It is an extensively utilized model considerably for making it simpler to translate and increment speculation by lessening fluctuation [19]. The relevant features in the dataset are complaints, age, harassment, and gender. LASSO feature selection method was used to select relevant features in the dataset that will be used for training.

3.3. Naïve Bayes Model

To classify the data for predicting if tenants will be harassed or not, the research made use of Naïve Bayes algorithm that is based on Bayes theorem. It is a probabilistic classifier, which means it predicts based on the probability of an object.

$$P(d/X) = P(X/d) * P(d)/P(X) \quad (1)$$

P(d): the probability of the hypothesis is true (regardless of the data).

P(X): the probability of the data being true (regardless of the hypothesis).

P(d/X): the probability of hypothesis d given the data X.

P(X/d): the probability of data X given that hypothesis d was true.

3.4. Logistic Regression

Logistic regression is a classification algorithm, used when the value of the target variable is categorical. Logistic Regression model can be written in a linear form as in equation 2:

$$In(\frac{P}{(1-P)=\beta_0}) + \beta_1 X_1 + \beta_2 X_2 + \cdots \quad (2)$$

Where:

P = Probability of Event, X_1, X_2… are the independent variable values.

The probability of event can be determined using equation

$$P = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots)}} \quad (3)$$

Where:

P = output between 0 and 1 (probability estimate)

e = base of natural log succinct as possible (for example, do not differentiate among departments of the same organization).

3.4. Support Vector Machine (SVM)

A Support Vector Machine or SVM is a machine learning algorithm that looks at data and sorts it into one of two categories.

Support Vector Machine is a supervised and linear Machine Learning algorithm most commonly used for solving classification problems and is also referred to as Support Vector Classification [20].

$$MINIMIZE_{a_{0,\dots,}a_m} : \sum_{j=1}^{n} MAX\,\{0, 1 - (\sum_{i=1}^{m} a_i x_{ij} + a_0) y_j\} + \lambda \sum_{i=1}^{m} (a_i)^2 \quad (4)$$

Where:

n = number of data points

m = number of attributes

x_{ij} = ith attribute of jth data point

y_j = 1 if data point is blue, -1 if data point is red.

4. Implementation and Results

The Jupyter notebook as a framework was used to run our python code to implement naïve Bayes, logistic regression and support vector machine on our data. Some libraries in python were used for the implementation such as pandas, NumPy, sklearn, and pickle. Pandas is a module in python used for data processing to view data. Numpy is a python machine learning library that helps to represent data in an array. Sklearn is a module in python where the libraries used for machine learning classifiers are stored. In the process of executing this work the following steps were taken:

- Import and load data
- Data encoding
- Training using Naïve Bayes
- Training using Logistic Regression

- Training using Support Vector Machine
- Evaluate model based on prediction

The comparative analysis of the result of the experiment is depicted in Table 1 and in Figure 1.

Table 1. Training Result

Classifier	Accuracy	Precision	Recall
Naive Bayes	94.7%	92%	98.8%
Logistic Regression	95.4%	93%	99%
Support Vector Machine	62%	59%	86%

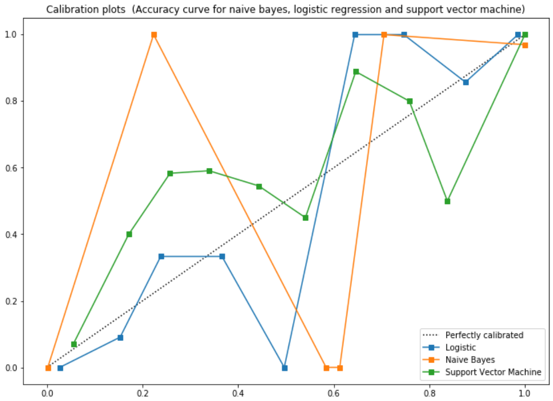

Figure 1. Calibration curve for naive Bayes, logistic regression and support vector machine

The support vector machine model was able to classify 62% of the input in-stances correctly. It exhibited a precision of 59% in average and recall of 86% while naïve bayes has an accuracy of 94.7%, a precision of 92% and recall of 98.8%, logistic regression has an accuracy of 95.4%, a precision of 93% and recall of 99%. The results show clearly that the logistic regression and naive bayes performs well compared to the support vector machine model.

5. Conclusions

With the increase in harassment, there is a huge demand for advanced systems and new approaches to better protect tenants. The development of a tenant harassment prediction system using Naïve Bayes, Support Vector Machine and Logistic Regression model has been carried out in this research. The system unites past harassment complaint records in particular zones and models such using Naïve Bayes classification, Support Vector Machine and Logistic regression to predict harassment. The approach is implemented using PYTHON JUPPYTER and performance is evaluated using accuracy, precision, and recall. The Logistic Regression model is compared with the Naïve Bayes model and Support Vector Machine model. The results obtained from the implementation of the model prediction, accuracy evaluation, and classifications proved that the Logistic Regression model approach to tenant harassment pre-diction performed better than the Naïve Bayes model with an accuracy of 95.4% and Support Vector Machine model with an accuracy of 62%.

6. References

[1] White, S. M. (2018). 10 Examples of Landlord Harassment and 10 Examples of Tenant Harassment. Retrieved from RentPrep: https://rentprep.com/property management/landlords-guideavoid-harassment/ (Access Date: 7 September 2022).

[2] Eberlin, E. (2020). Landlord Actions That Are Considered Harassment. Retrieved from the balance small business: https://www.thebalancesmb.com/is-yourlandlord-guilty-of-harassment-4125876 (Access Date: 16 September 2022).

[3] Michael Ayodele Olukolajo, Matthew Taiwo Ogun-gbenro and Amos O. Adewusi (2015). Tenants' Characteristics and Rent Default Tendencies in Akure Residential Property Market, African Journal of Built Environment Research, Vol. 2, No. 1, July 2018, 41-54.

[4] Soludo, C. (2007). Nigeria's Financial System Strategy 2020 Plan "Our Dream". FSS 2020 International Conference. Abuja Nigeria.

[5] Jothi, N., Rashid, N. A., and Husain, W. (2015). Data mining in healthcare - A Review. Procedia Computer Science, 72, 306–313.

[6] Rebecca, A., and Johnson, T. Y. (2019). Using Machine Learning to Help Vulnerable Tenants in New York City. In ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '19), July 3–5, 2019, Accra, Ghana. 11 pages. DOI: 10.1145/3314344.3332484.

[7] Gürcan, F. (2018). Multi-Class Classification of Turkish Texts with Machine Learning Algorithms," 2018 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). pp. 1-5, DOI: 10.1109/ISMSIT.2018.8567307.

[8] Kumar, A., Mangla, S.K., Luthra, S., Rana, N.P. and Dwivedi, Y.K. (2018). Predicting changing pattern: building model for consumer decision making in digital market. Journal of Enterprise Information Management, Vol. 31 No. 5, pp. 674-703. DOI: 10.1108/JEIM-01-2018-0003.

[9] Riyanto Jayadi, H. M. (2019). Employee Performance Prediction using Naïve Bayes. International Journal of Advanced Trends in Computer Science and Engineering, 8, 3031-3035. DOI: 10.30534/ijatcse/2019/59862019.

[10] Palaniappan, S., and Awang, R. (2008). Intelligent Heart Disease Prediction System Using Data Mining

Techniques" IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.8, August.

[11] Ye, T., Johnson, R., Fu, S., Copeny, J., Donnelly, B., Freeman, A., Lima, M., Walsh, J., and Ghani, R. (2019). Using Machine Learning to Help Vulnerable Tenants in New York City. In ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '19), July 3–5. Accra, Ghana. 11 pages. DOI: 10.1145/3314344.3 332484.

[12] Molina-Gil, J. A. S. G. (2019). Harassment Detection Using Machine Learning and Fuzzy Logic Techniques. DOI: proceedings2019031027.

[13] Masoom Patel, P. S. (2020). Bully identification with machine learning algorithms. Journal of Critical Reviews ISSN-2394-5125 Vol 7, Issue 6, 2020. DOI: 10.31838/jcr.07.06.74.

[14] Yu, C-H., Ward, M. W., Morabito, Me., and Ding, W. (2011). [IEEE 2011 IEEE International Conference on Data Mining Workshops (ICDMW) - Vancouver, BC, Canada (2011.12.11-2011.12.11)] 2011 IEEE 11th International Conference on Data Mining Workshops - Crime Forecasting Using Data Mining Techniques. 779–786. DOI: 10.1109/icdmw.2011.56.

[15] Md. Milon, I., Hasib, I., Md. Rezwanul, H., and Md. Kamrul, H. (2017). [IEEE 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) - Dhaka, Bangladesh (2017.12.21-2017.12.23)] 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) - Prediction of breast cancer using support vector machine and K-Nearest neighbors. 226–229. DOI: 10.1109/R10-HTC.2017.8288944.

[16] Li, X., Lord, D., Zhang, Y., and Xie, Y. (2008). Predicting motor vehicle crashes using Support Vector Machine models. 40(4), 1611–1618. DOI: 10.1016/j.aap.2008.04.0 10.

[17] Anuja, V., and Kumari, R. (2013). Classification Of Diabetes Disease Using Support Vector Machine. International Journal of Engineering Research and Applications, 1797-1801.

[18] Axelberg, P.G.V.; Irene Yu-Hua Gu; Bollen, M.H.J. (2007). Support Vector Machine for Classification of Voltage Disturbances. 22(3), 0–1303. DOI: 10.1109/tpwrd.2007.900065.

[19] Gnaneswara Rao Nitta, B. Y. (2018). LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type. Service-Oriented Computing and Applications. DOI: 10.1007/s11761-018-0251-3.

[20] Radhika. (2020, October 23). The Mathematics Behind Support Vector Machine Algorithm (SVM). https://www.analyticsvidhya.com/blog/2020/10/the-mathematics-behind-svm/ (Access Date: 21 September 2022).