Troubled Property Analysis Using Machine Learning

Daniel Zhao, Gabriel Radich
Catholic University of America
CSC 484 - Spring 2024
zhaop@cua.edu, radichg@cua.edu

Abstract

This project focuses on developing a Troubled Property Analysis course specifically tailored for stakeholders interested in real estate investment in Montgomery County, Maryland. Utilizing machine learning methodologies, the course aims to equip participants with the knowledge to assess distressed properties effectively. The course curriculum will cover techniques for identifying troubled properties, analyzing market trends, and making informed investment decisions. Two key datasets will be utilized: monthly reports on troubled properties in Montgomery County and housing market data from reputable sources. This analysis will employ various machine learning algorithms, including and limited to K-nearest neighbors, Decision Tree, Nearest Centroid Classifier and Multilayer Perceptron, to analyze and interpret the data. Python will be the only programming language, with libraries such as scikit-learn and pandas used for implementation. By providing actionable insights and practical skills, this project seeks to empower participants in navigating the complexities of troubled property investment in Montgomery County.

1 Introduction

Real estate investment presents lucrative opportunities but also involves inherent risks, particularly when dealing with troubled properties. In Montgomery County, Maryland, stakeholders face the challenge of identifying and analyzing distressed properties amidst a dynamic real estate market. This project endeavors to address this challenge by developing a comprehensive Troubled Property Analysis course tailored specifically for Montgomery County.

The course aims to equip participants with the knowledge and skills necessary to effectively identify, assess, and invest in troubled properties in the county. By leveraging machine learning techniques and utilizing relevant datasets, the course will provide actionable insights into market trends, property conditions, and investment potential.

Through a combination of theoretical concepts and practical applications, participants will gain a deep understanding of the factors contributing to property distress, strategies for mitigating risks, and opportunities for maximizing returns. Moreover, the course will foster critical thinking and decision-making skills essential for navigating the complexities of real estate investment in Montgomery County.

With a focus on practicality and relevance, this project seeks to empower stakeholders with the tools and knowledge needed to succeed in troubled property investment, ultimately contributing to the growth and sustainability of the real estate market in Montgomery County, Maryland.

2 Data

The dataset utilized in this project comprises housing code inspection results for multifamily rental properties in Montgomery County, Maryland, conducted in accordance with Executive Regulation 05-17. This regulation outlines criteria for designating properties as "Troubled," "At-Risk," or "Compliant" based on the severity and quantity of health and safety violations identified during inspections.

Key features of the dataset include:

Classification of Properties: Properties are classified as "Troubled," "At-Risk," or "Compliant" based on the severity and quantity of housing code violations identified during inspections. A fourth label “TBD” is assigned to unclassified housing complexes. A "Troubled Property" typically exhibits numerous and/or severe violations, necessitating corrective action plans and at least annual inspections.

Inspection Results: The dataset provides detailed information on the outcomes of housing code inspections conducted on multifamily rental properties. This includes specific violations identified, severity scores assigned to violations, and overall compliance status.

Annual Updates: The dataset is updated annually to reflect inspection results for properties inspected within a given fiscal year, typically spanning from July 1st to June 30th.

Predictive Insights: By analyzing this dataset, stakeholders can gain insights into trends and patterns related to housing code violations, identify troubled properties requiring

intervention, and assess the effectiveness of corrective action plans implemented by property owners.

By leveraging this dataset, stakeholders, including policymakers, housing authorities, property owners, and community organizations, can work collaboratively to improve housing quality, ensure tenant safety, and address housing code violations effectively in Montgomery County. Additionally, the dataset serves as a valuable resource for monitoring compliance with regulatory standards and informing targeted interventions to mitigate risks associated with troubled properties.

3 Related Work

Prior research and initiatives have addressed various aspects of real estate investment and property analysis, laying the groundwork for this project's objectives. Several studies have explored the use of machine learning algorithms for predicting housing market trends, property valuation, and investment decision-making. For instance, research by Li and Yu (2018) demonstrated the effectiveness of machine learning models in forecasting housing market fluctuations and identifying profitable investment opportunities.

Furthermore, courses and training programs focusing on real estate investment and property analysis have been developed to educate stakeholders on best practices and strategies. While many existing courses provide valuable insights into general real estate principles, there is a notable gap in specialized training tailored specifically for troubled property analysis in specific regions, such as Montgomery County, Maryland.

In the context of Montgomery County, initiatives aimed at addressing distressed properties have been undertaken by local government agencies and community organizations. These efforts often involve data collection, analysis, and intervention strategies to revitalize neighborhoods and mitigate the negative impacts of troubled properties on communities.

However, despite these efforts, there remains a need for a comprehensive course dedicated to troubled property analysis in Montgomery County, integrating both theoretical concepts and practical applications. This project seeks to fill this gap by leveraging machine learning techniques, relevant datasets, and specialized curriculum development to empower stakeholders with the knowledge and skills necessary for successful troubled property investment in the county.

4 Models

Four models were implemented; the following sections describe each in detail.

4.1 Decision Tree Classifier

This model uses binary splitting based on information gain to split the data into classes. The decision tree we trained used Gini impurity as the splitting criterion. Gini calculates, based on probability, the accuracy of a given split. If the Gini impurity is 1, then the classes overlap completely after the split. If the Gini was 0, the tree separated the data into two mutually exclusive classes. Obviously, for classification, the lowest Gini impurity is desired. Gini is calculated according to the following formula.

(1)

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

In this formula, p(i) is the probability of a data point being assigned a certain class label, and 1–p(i) is the probability of a data point being assigned any other class label. C is the total number of class labels.

Using a decision tree we were able to classify the data with 99.18% testing accuracy and 100.0% training accuracy. This was obtained with the out-of-the-box DecisionTreeClassifier model from the sklearn.tree module. The “accuracy_score” function and RMSE were used to evaluate the performance of the decision tree.

4.2 Nearest Centroid Classifier

This model randomly selects ‘centers’ in the data, one for each class. During the training, the classifier moves each center to the true center of each class. Once the centers stop moving, the training stops. For this classifier,the euclidean method was chosen to calculate the distance from each center to the other data points. The euclidean distance function is

(2)

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

In this formula, p and q are two distinct rows of a dataset, and the distance between them is calculated based on their values.

4.3 K-Nearest Neighbors Classifier

This is the simplest model one can use to analyze data, but it is powerful nonetheless. This model uses the euclidean distance, defined in formula (2) to find the distance between each row and its k nearest neighbors. Then, the classifier takes the majority votes of the targets of the neighboring rows and assigns the class label corresponding to the highest vote. The best k value for our dataset is k = 6. This was calculated using RMSE over several iterations of k.

4.4 MLP Classifier

This model optimizes the cross-entropy function. This model uses a solver and an activation function. For the activation functions we used ReLu, tanh and identity. The cross-entropy function is determined by the following formula:

(4)

$$L_{\text{cross-entropy}}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{i} y_i \log(\hat{y}_i)$$

In this loss function, y is the true target vector, and y_hat is the predicted target vector.

5 Implementation

All code was written in Python 3, using Jupyter Notebook 6.5.4. and Google Colab.

All of the models used in this analysis are from Scikit-Learn. Experiments were run on local CPU-only.

6 Results and Discussion

The most accurate model used was the decision tree, which classified the data with a 99.18% testing accuracy and a 100.0% training accuracy (See Table 1). This high accuracy was certainly not expected, because the data points were taken from distinct rental housing complexes in the Montgomery County area. One would think that two different complexes would not necessarily have anything in common, according to the features in the dataset. A few reasons for this extremely high accuracy could be attributed to a pattern of some kind that the data follows naturally, which is akin to the decision tree's algorithm. Another possibility is that some of the features of the dataset are generated by formulas applied to

other columns in the dataset. A third possibility is that the decision tree was the best model for classifying the data. We found this to be the most puzzling result.

The second model used was the nearest centroid classifier, which had a final testing accuracy of 51.61% (Table 1). So, this model was the least accurate of the four we trained.

The third model used was a k-nearest neighbors classifier. The k parameter resulting in the lowest RMSE was k = 6. The final testing accuracy with k = 6 was 80.65% (Table 1).

The fourth and final model we trained was the multi-layer perceptron. We trained several models with different activation functions. The three activation functions we used were ReLu, tanh() and identity. We iterated the model using 20 different hidden layer sizes ranging from n = 100 to n = 1000 in increments of 50. Ultimately, we found that the most accurate MLP classifier classified the data with a 90.32% testing accuracy (Table 1). This was the classifier, which employed the tanh() activation function.

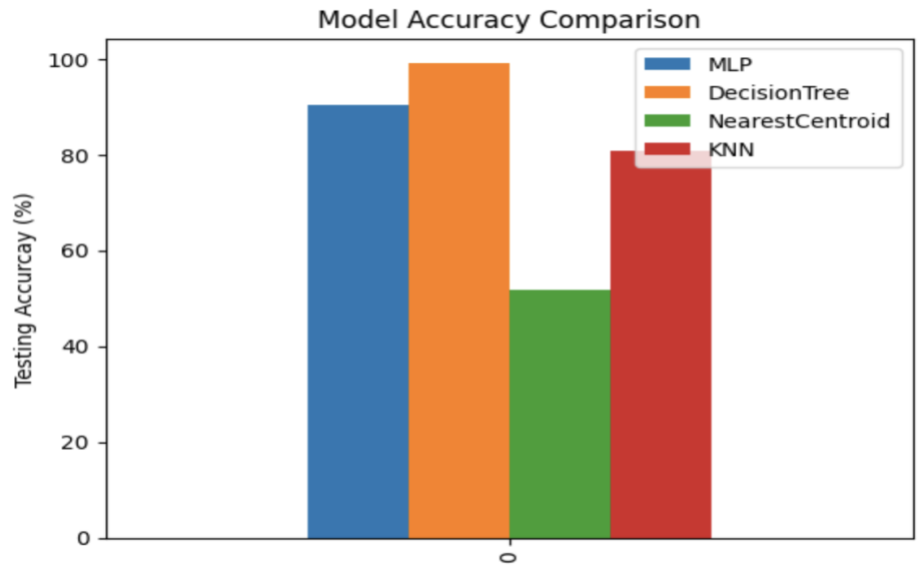
In conclusion, the two most accurate models were the Multilayer Perceptron and the Decision Tree. No other models came within 10 percent of their testing accuracy. These two models could be used to classify new complexes, using inspection reports from the county.

Table 1: Testing Accuracy and Parameters for the Classifiers

Model	Testing Accuracy* (%)	Parameters
MLP Classifier	90.32	150 hidden layers and tanh() as the activation function.
Decision Tree	99.18	
Nearest Centroid	51.61	—
K-Nearest Neighbors	80.65	k = 6

*All these values were computed using the “accuracy_score” function from Scikit-Learn.

Figure 1: Graph of Accuracy Across All Classifiers



7 Future Directions

Enhanced Predictive Modeling: Incorporating advanced machine learning algorithms and predictive analytics techniques can improve the accuracy and reliability of identifying troubled properties and assessing housing code violations. Techniques such as ensemble learning, anomaly detection, and natural language processing can be explored to enhance the predictive capabilities of the model.

Real-Time Monitoring and Intervention: Developing a real-time monitoring system that integrates data from ongoing housing inspections, property maintenance reports, and tenant complaints can enable proactive identification of troubled properties and timely intervention to address emerging issues. This would require the implementation of robust data collection, integration, and analytics infrastructure.

Community Engagement and Empowerment: Engaging with local communities, tenant associations, and advocacy groups can foster collaboration and empower residents to report housing code violations, advocate for their rights, and participate in decision-making

processes related to housing quality and safety. Implementing outreach programs and digital platforms for community feedback and participation can facilitate greater transparency and accountability in housing management practices.

Holistic Property Rehabilitation Strategies: Moving beyond reactive measures, there is a need to develop holistic rehabilitation strategies for troubled properties that address underlying structural, economic, and social factors contributing to housing code violations. This may involve leveraging public-private partnerships, accessing funding sources for property rehabilitation, and implementing supportive services for vulnerable tenants.

Policy and Regulatory Reforms: Continuously reviewing and updating housing regulations, enforcement mechanisms, and inspection protocols can ensure their alignment with evolving housing market dynamics, emerging risks, and community needs. This may include streamlining inspection processes, strengthening penalties for non-compliance, and implementing incentives for property owners to maintain high standards of housing quality.

Data Sharing and Collaboration: Promoting data sharing and collaboration among government agencies, housing providers, researchers, and community stakeholders can facilitate the exchange of best practices, data insights, and resources for addressing housing code violations and improving housing conditions. Developing data-sharing agreements, interoperable systems, and collaborative platforms can facilitate knowledge exchange and collective action towards shared housing goals.

By pursuing these future directions, stakeholders can contribute to the creation of safe, healthy, and sustainable housing environments that meet the needs and aspirations of all residents in Montgomery County, Maryland.

Acknowledgments

Many thanks to Dr. Chaofan Sun, who advised us on this project. Dr. Sun and Vy Nguyen, the TA of CSC 484, deserve high praise for making the course highly engaging, despite the logistical challenges posed by the sheer size of the class.

References

[1] Yu SL, Li Z (2018) Forecasting Stock Price Index Volatility with LSTM Deep Neural Network, In: Tavana M, Patnaik S. Editors, Recent Developments in Data Science and Business Analytics, Cham: Springer Nature, 265–272.