Predicting Property Leads for Real Estate Wholesaling

Tamar S Tenenbaum

CU-400: Research Writing

Dr. Daniels

July 29, 2024

Predicting Property Leads for Real Estate Wholesaling

Abstract

The goal of this study was to create a predictive model for real estate lead generation. The project began with an extensive review of existing literature, providing a strong theoretical foundation. Analysis and preprocessing followed, in order to ensure the data was suitable for model creation, and to offer insights into how the company should interpret trends in the data when evaluating potential real estate purchases. Data preparation included normalizing values, removing features, and oversampling. Once done, a series of XGBoost models with different hyperparameter configurations and varying test-train splits were developed to predict lead generation potential. The performance of each model was evaluated, concluding in the most effective model. The model with the best performance had a test split of .2, and used deep trees with numerous boosting rounds and a slow learning rate, resulting in a highly complex and effective predictive tool. Its success underscores the importance of a detailed and nuanced approach to this data analysis, as it was able to capture intricate patterns and subtle variations within the data. Despite its strengths in accuracy and recall, the model would benefit from adjustments to minimize false positives, enhancing its utility in narrowing down potential leads for real estate wholesalers. This study represents an initial yet crucial step in creating a robust predictive model. The findings indicate that by implementing these models, the studied real estate wholesaling company can significantly improve their customer targeting strategies. While further refinement is needed, the groundwork laid in this research is promising and sets the stage for substantial improvements in identifying and pursuing potential leads, ultimately transforming the real estate wholesaling process.

Introduction

Real estate wholesaling could be a challenge, as the target was to buy homes that are not on the market. Although the industry yielded good margins, the low success rate had potential to be aided. This research aimed to assist the prediction of good leads for a particular wholesaling real estate company. I Explored key areas relevant to developing a predictive model for real estate leads, focusing on data preprocessing, handling imbalanced data, and various machine learning techniques in the field in order to give light to how to approach this task. Specifically, the goal of this study was to predict whether a property would be a good lead based on available property data.

The process of buying and selling property is unlike the transaction of any other good and requires far more customer thought and involvement. (Xiuzhi 2022). Studies indicated that the methods used to target homeowners and the characteristics of the individuals themselves significantly influenced the quality of leads in real estate. For example, studies found that online advertisements positively impacted new home sales, especially in lower-tier cities and among higher-income residents, suggesting that targeted digital marketing can enhance lead quality (Xiuzhi, Z., Zhang, Y., & Zhijie, L. (2022). Moreover, key factors such as property cost,

location, and interior decoration were crucial in influencing real estate prices, which could also be extended to understanding lead quality (Huang-Mei et al., 2021). Sources also said it was best to target consumers on factors such as age, sex (Sknarev, 2020), the property area's market (Sirgy, n.d., pp. 3-6), the consumer's expectations on profit (Case, 19088), and people with reason or intention to move (Glower, M., Haurin, D. R., & Hendershott, 1997). However, while these insights are valuable, much information on the homeowner themselves was not available in this research. Consequently, this limitation restricted a more granular analysis of how these factors specifically affected lead quality within the dataset at hand

Data in real estate posed a particular challenge. Between datasets in real estate, the levels of observation are often inconsistent, identifiers usually don't match, and there can be a lack of standardization. To ensure the quality and reliability of the dataset, one must leverage effective data cleaning strategies. The data preparation process within real estate is different for every case. The approach taken depends heavily on the goal of the study and quality of the data used (Krause & Lipscomb, 2016).

As imbalanced datasets are a common issue in real estate, it was vital to address the challenges of learning from imbalanced data. Effective approaches were the tweaking of algorithms to assign higher costs to less represented objects and oversampling, helping achieve a balanced performance across different classes (Krawczyk 2016). Haixiang et al. (2017) further explored methods for dealing with imbalanced data, such as preprocessing and cost-sensitive learning.

The development of decision support systems (DSS) for real estate was explored in previous studies. These systems aimed to automate the buying and selling process by evaluating various criteria important to buyers. For instance, a DSS developed for real estate transactions considered factors such as location, price, and buyer preferences (Zavadskas, Kaklauskas, & Banaitis, 2010). While the goal of full automation remained a challenge, these systems provided valuable insights into the factors that drive successful transactions.

Other sorts of predictive modeling have been used to identify real estate opportunities and assess property values. For example, models have been created to predict real estate prices using machine learning algorithms such as support vector regression, k-nearest neighbors, ensembles of regression trees, and multi-layer perceptrons (Baldominos et al., 2018), Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Fuzzy Least-Squares Regression (FLSR) (Sarip, Hafez, & Daud, 2016). These models highlight the importance of incorporating a large set of features to achieve accurate predictions. In China, decision trees and random forest models have been used to demonstrate factors that significantly impacted real estate prices, and achieved high accuracy in predictions (Huang-Mei, Chan, Jia-Ying, Xue-Qing, & Zne-Jung, 2021). One can apply the underlying principles of research focusing on price prediction to predicting the likelihood of a property being a good lead.

This study hypothesizes that it is possible to create a model that can predict the likelihood of a property being a good lead for a specific company, while only using basic attainable property data. Previous research that demonstrated the effectiveness of machine learning models

in predicting real estate prices and the importance of various property attributes supported this endeavor.

Methods

The goal of the model was to predict if a property will produce a response from the owner, be it positive or negative, thereby being a good lead. Such a model could increase the success rate of buying properties to filter lists for more targeted results.

Before leveraging the data to predict a property's label as a lead or not, a cleansing of the data and preprocessing was absolutely critical. This was done using several libraries in python such as numpy, pandas, matplotlib, seaborn, fuzzy-wuzzy, and rapid fuzz. Once the data was ready, a model was chosen for implementation.

This study employed the XGBoost model, known for its performance in handling imbalanced datasets with many features and missing values. XGBoost is a gradient boosting framework that uses tree-based learning algorithms, which are particularly effective for binary classification tasks such as this one. The model's advanced boosting and regularization techniques mitigate the effects of class imbalance by appropriately weighting less frequent classes, leading to improved accuracy and performance metrics (Chen & Guestrin, 2016).

The initial features used in the model were 'Property City', 'Property State', 'Property Zip', 'Mailing City', 'Mailing State', 'Mailing Zip', 'Bedroom', 'Bathroom', 'Apporx Sqft',' Lot Size Sqft', 'Effective Year Built', 'Tax Assessed Value', and 'Last Sold Price'. These were used to predict the column 'Lead at all'. The one-hot-encoding from side-kick-learn was applied to all categorical features. These features were chosen because they were available to the public and did not require any purchases. The use of free features gave way to the most profitable model, as more data would not have to be acquired.

Using side-kick-learn's train_test_split, two test splits were done: 50%-50%, and 80%-20%. Within each split, 4 models were created. The first 3 took the duplicated test split and performed XGBoosts with different hyperparameters. A gridsearch and k-fold would have been beneficial in this context, but unfortunately the computational power was unavailable.

Model 1 was a basic model with a moderate approach, featuring n_estimators=100, max_depth=4, learning_rate=0.1, subsample=0.8, and colsample_bytree=0.8. 100 boosting rounds indicated a simpler model that could have lower variance but higher bias. A max depth of 4 allowed the model to capture interactions up to four levels deep which balanced the ability to learn complex patterns while minimizing the risk of overfitting. The learning rate of 0.1 caused a steady progress in learning, ensuring that each step contributes to model improvement. Subsampling 80% introduced randomness, which helped prevent overfitting, while using 80% of the features at each split reduced the model's complexity.

In contrast, Model 2 was designed to capture more complex interactions and provided a more granular learning process to adjust for possible underfitting. It used max_depth=6, min_child_weight=1, gamma=0, subsample=0.8, colsample_bytree=0.8, learning_rate=0.05, and n_estimators=200. This combination ensured that each step was smaller and more precise than

Model 1, while the larger number of boosting rounds compensated by allowing more iterations to achieve robust learning. The subsample and colsample_bytree parameters remain consistent with Model 1, maintaining the approach of using 80% of the data and features to stabilize learning.

Model 3 used regularization measures to prevent overfitting, with n_estimators=200, max_depth=3, learning_rate=0.05, subsample=0.7, colsample_bytree=0.7, reg_alpha=0.1, and reg_lambda=1.0. The max depth was reduced to 3 which made the trees shallower and less complex, but vulnerable to missing deeper data patterns. The learning rate and number of estimators were the same as in Model 2, ensuring a consistent learning process through a higher number of smaller steps. Lowering the subsample to 70% and colsample_bytree to 70% introduced more randomness, which helped in reducing overfitting but could potentially lower accuracy. Furthermore, the addition of reg_alpha=0.1 (L1 regularization) and reg_lambda=1.0 (L2 regularization) imposed penalties on the model's complexity.

To summarize these three models, the hyperparameters each reflected different strategic choices. Model 1 took the middle ground between complexity and generalization, Model 2 aimed to capture deeper and more intricate patterns with increased learning iterations, and Model 3 focused on robust regularization to prevent overfitting.

The strategy used by the wholesaling company was generally to 'cast a wide net' in property hunting. Meaning, the company gathered thousands of properties almost randomly and marketed to each one at a low cost. The profit made from the bought properties strongly outweighed the loss of marketing to dead-end properties. Therefore, a property falsely predicted negative was exponentially more detrimental than a false positive. Because of this, one of the most important metrics in this research was recall, which was the accurately predicted positives out of all positives.

In total 8 similar models were implemented to compare results. It was important to note that even after oversampling, the accuracy metric may have still been very high due to the greater majority of properties not producing a lead. This was acceptable, as the focus was on models that performed best on the finer details. The true measure of each models' effectiveness lay in its ability to accurately predict across both classes, as well as the avoiding of false negatives.

Data Overview and Preparation

The research required use of multiple datasets. A sequence of events took place before the data was attainable for analysis and modeling. The wholesaling real estate company first went on a property website and chose thousands of potential sellers filtered based on intuition. I acquired many of these lists and referred to them as the raw datasets. The company then bought the property owners' phone numbers and marketed to them through ringless voicemails, cold calling, and SMS blasts. They also marketed through youtube, billboards, and other means. If a property owner responded, the property data was added to a platform for the sales representatives. This platform served to keep track of leads and ensured a process was followed by the sales representatives. The status of a lead on this platform constantly revolved as a sale

progressed and regressed. I downloaded the data from this platform at random 4 times over 2 months. This is the lead dataset. The lead dataset and the raw dataset are combined for model creation.

After hours spent preprocessing the data, the final data set had 129673 rows and 28 columns. The columns of the final data set were 'Lead Status', 'Score', 'Lead Quality', 'Lead at all', 'Lead Source', 'Campaign Name', 'First Name', 'Last Name', 'Gender', 'Phone Area Code', 'Phone Number', 'Property Street Address', 'Property City', 'Property State','Property Zip', 'Mailing Street Address', 'Mailing City', 'Mailing State', 'Mailing Zip', 'Bedroom', 'Bathroom', 'Apporx Sqft', 'Lot Size Sqft', 'Effective Year Built', 'Tax Assessed Value', 'Last Sold Price', and 'Lead Created Date'.

The original lead data had sixty three columns and 6816 rows. Twenty four of those columns were empty. Of those which contained data, fifteen were filled only when that column's lead progressed. These fifteen columns were unnecessary, since the focus was on predicting the quality of a lead, and these columns were connected to the target. After removing empty and unnecessary columns, breaking addresses down, and adding a few columns, the final dataset contained thirty one columns.

Due to the nature of the data, many properties were repeated throughout the lead datasets, but were labeled differently each time. I located all duplicates and deleted ones where the lead was worse than its match. This allowed the 'Lead Quality' column to classify each property based on its best status. Part of his project was the analysis of the probability that a lead was ever good, not if it would remain good, or if it was not good to begin with. Therefore the optimal approach was not to have repeat samples. After doing so, the final lead data set had 1734 rows, and 31 columns.

Within the lead data set there were many null variables. Duplicates or overlapping property addresses between the raw and lead datasets were used to fill null values in the lead datasets. Unfortunately there were not many overlaps and many values remained empty. Once the matches were used to fill missing variables in the lead dataset, the two were combined and duplicates were removed.

As a property progressed through the system, the lead status received one of nine labels. The target variable was a column named 'Lead at all', which was created from the existing 'Lead Status' column. Additionally, a column 'Lead Quality' was developed, it grouped the labels from 'Lead Score' for broader analysis. A 'Bad' lead quality had the lead status "Dead Lead" or "Warm Lead". An 'Ok' lead was labeled "New Leads", "No Contact Made", and "Contact Made". A good lead was labeled "Hot Leads🔥", "Appointments Set", "Offers Made", and "Under Contract". Properties in the raw data that did not result in a lead were labeled as "None" in this column.

The columns 'Score', 'Lead Quality', and "Lead at all' were derived from 'Lead Status'. 'Score' was the numerical representation of 'Lead Status', and 'Lead Quality' was a grouping of the 'Lead Status' as discussed above. 'Lead at all' was a binary target of 0 and 1, 0 being the property did not produce any sort of lead, and 1 being it did. The properties that resulted in a lead

were 0.96% of the data. These values were heavily skewed as most properties did not result in a lead, posing a challenge when creating a model.

'Lead Source', 'Lead Created Date', and 'Campaign Name' were where and when the lead came in from. The vast majority of leads were produced from cold calling, and the only leads that resulted in a ‘Good’ lead came from cold calling and sms blasts (Fig. 1). Rows of no lead were filled with ‘None’. Since these values were directly tied with my target, these columns were inapplicable in my model.

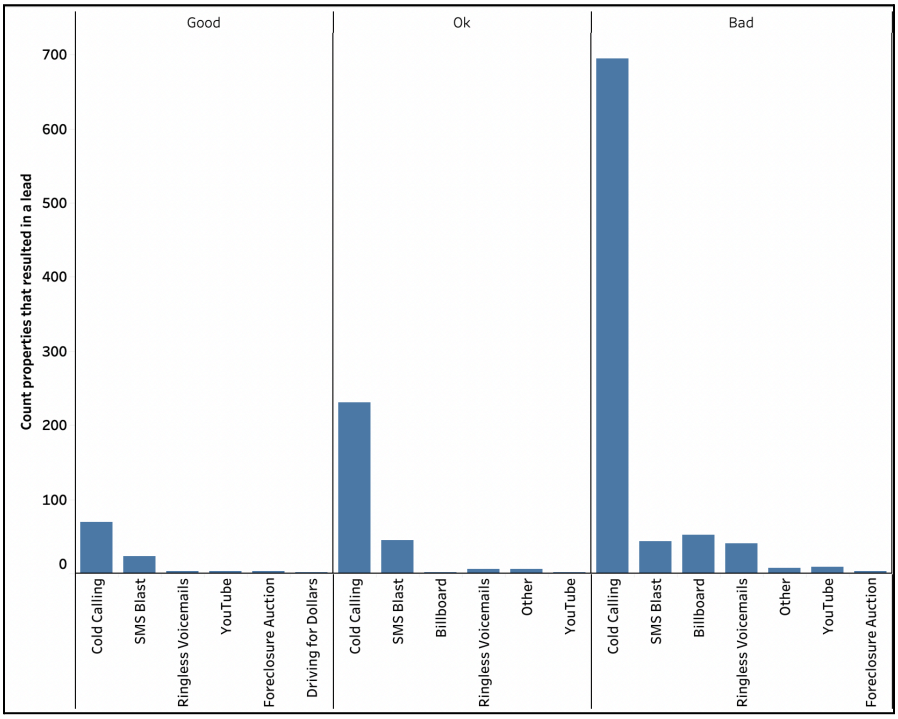

Fig. 1.

First Name and Last name columns had 17,511 and 108 null values respectively. With no intention to use these columns for model creation, nulls were irrelevant. From the first name, the gender-guesser library was used to create the column ‘Gender’. This column contained 4 options: ‘male’, ‘female’, ‘andy’, meaning the name was androgynous, ‘mostly_male’, ‘mostly_female’, meaning the name was mostly used by that gender but not exclusively, and ‘unknown’, meaning the name's gender was unknown, or there was no recorded name.

A chi-square test provided insight into the possible relationship between gender and lead. The results showed the chi-square statistic was 37.7235, which quantified the difference between the observed and expected frequencies. The p-value was extremely small (4.2878e-07), indicating that the probability of observing such a difference by random chance was almost impossible. With a p-value far below the common significance threshold, commonly 0.05, suggesting an association between gender and production of a lead. The vast majority of leads

came from male classified first names, at 52.62%, followed by female names at 25.98%, suggesting properties owned by males have a higher probability of being a lead.

The 'Phone Number' column was used to derive the ‘phone area code’. There were 125 total area codes within the data. The phone number was only available for analysis for properties where a lead was produced, and would therefore not be a good predictor. However, it was still useful for deriving insights to the data and future marketing. In a discussion with the company's CEO, he related that the phone number used for ringless voicemails had an area code 754, the area codes for cold calling were 786 and 954, and the numbers for sms blasts began with 954 and 561. One can hypothesize that more leads were produced when the property owner had the same area code as the marketing number because the number would appear familiar.

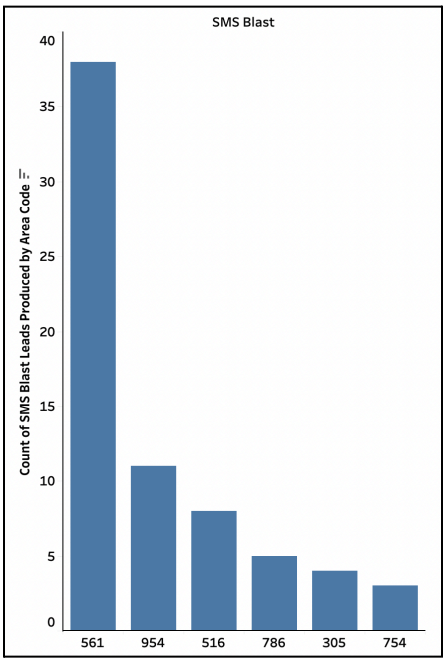

Fig. 2

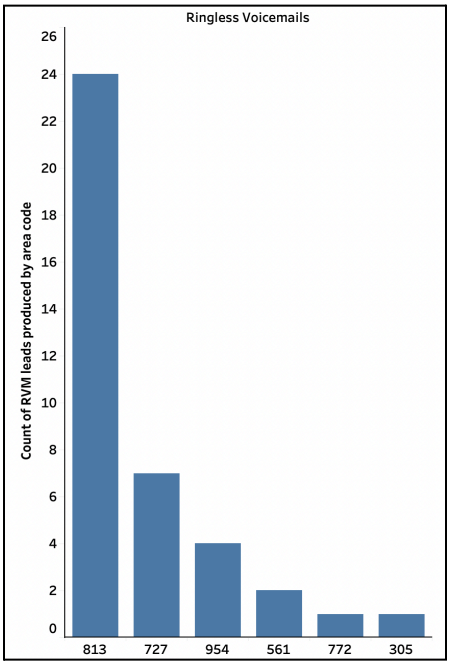

Fig. 3

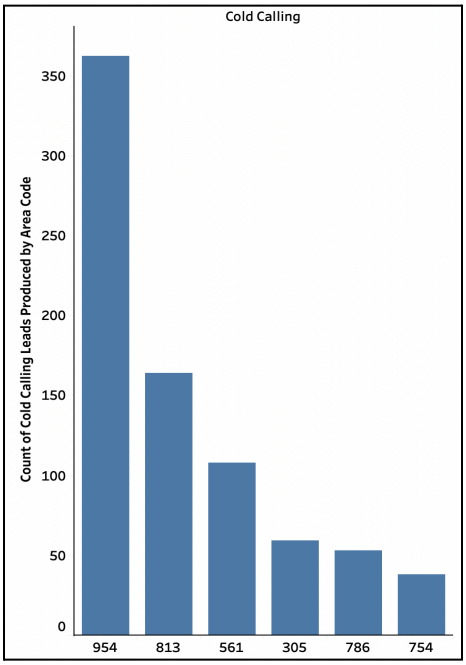

Fig. 4

The above figures display the top six area codes that resulted in the most leads for sms blasts, ringless voicemails and cold calling. A correlation could be seen in the numbers used for SMS and cold calling (Fig. 2 and Fig. 3). The most leads were produced by the area codes 561 and 954 for SMS blasts, and 954 and 786 were both at the top six for cold calling. There seemed to be no correlation between the phone number used for ringless voicemails and the resulting leads. It was important to note that there was no data on this topic on the area codes that did not result in any sort of lead. Therefore one could not conclude with just this whether the area code made a significant difference, yet the observation was noteworthy and should be developed.

From the property and mailing addresses were derived the street addresses, cities, and zip codes. Each contained null values ranging from 1 to 461. Using a proportional symbol map showed which property zip codes produced the most leads (Fig. 5). In addition, plotting the leads and the none-leads showed some consistency between the zip code and the possibility of a lead.

Fig. 6 showed the properties that did not result in a lead, and Fig. 3 showed those that did. The properties that didn't result in a lead were in concentrated regions where Florida is highly populated. Those that did result in a lead were dense in those areas as well, but also existed outside of them as seen in Fig. 7.

The 'Mailing City' and 'Property City' contained duplicate values written slightly differently. Many had a space before or after the city name, or were capitalized differently. Implementing panda's fuzzy-wuzzy library fixed this issue, and resulted in clean columns that could be analyzed. The heat map in Fig. 8 showed the cities that had the highest percentage of leads. This heat map excluded skews where only one specific property is targeted in a city. The city with the highest percentage of leads out of total inquiries was Lakeland, followed by Tequesta, Pembroke Park, and Hypoluxo.

Property City1	Didnt result in ..	Resulted in a le..
Lakeland	33.33%	66.67%
Tequesta	50.00%	50.00%
Pembroke Park	50.00%	50.00%
Hypoluxo	50.00%	50.00%
Palm Beach	80.00%	20.00%
Dania Beach	89.86%	10.14%
Clewiston	91.67%	8.33%
Juno Beach	93.33%	6.67%
Wilton Manors	96.30%	3.70%
Hallandale Beach	96.40%	3.60%
Margate	96.76%	3.24%
Lauderhill	97.68%	2.32%
Hollywood	97.74%	2.26%
Fort Lauderdale	97.86%	2.14%
Southwest Ranches	97.87%	2.13%
Pompano Beach	97.91%	2.09%
Greenacres	98.01%	1.99%
Lauderdale Lakes	98.08%	1.92%
Delray Beach	98.27%	1.73%
Tamarac	98.29%	1.71%
West Park	98.42%	1.58%
Deerfield Beach	98.63%	1.37%

Fig. 8

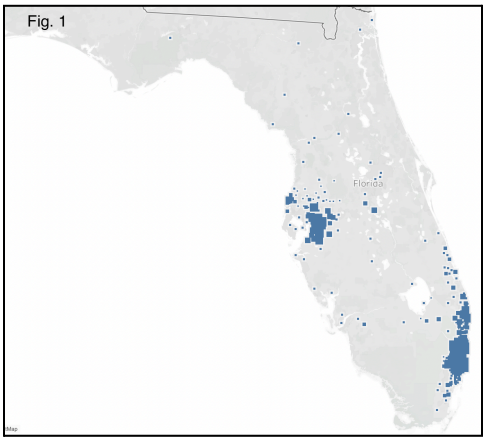

Fig. 5

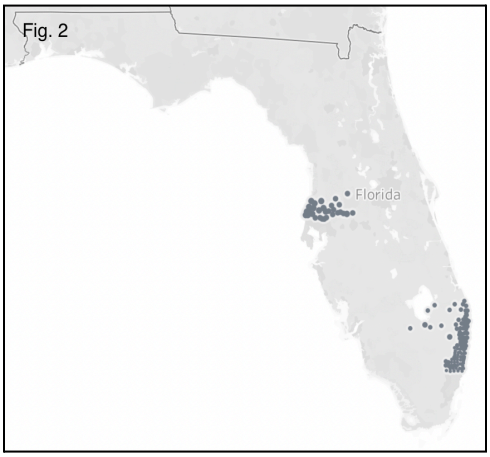

Fig. 6

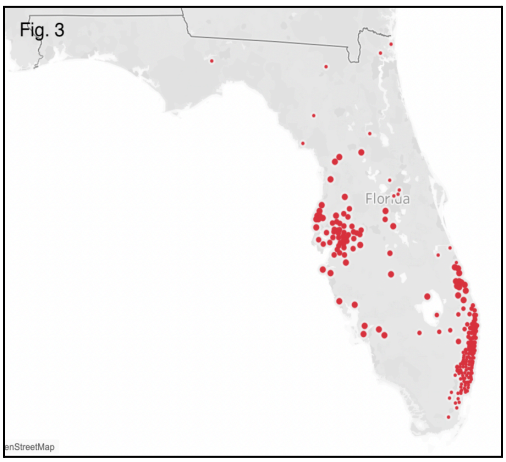

Fig. 7

The 'Bedroom' and 'Bathroom' columns represented the amount of bedrooms and bathrooms there were in each property. There were 38,121 null bedrooms and 6,763 null bathrooms. In addition, each had significant outliers coming from the raw data sets. Bedroom outliers that were greater than 65, and bathroom outliers greater than 54 were removed. These numbers were chosen because the large counts less than 65 and 54 belonged to multi-family properties. These box plots represented the distribution of bedrooms and bathrooms in properties to resulting leads (Fig. 9 and Fig. 10). The means of those resulting in a lead are significantly lower than those that don't.

The column 'Approx Sqft' had 396 missing values, and 'Lot Size Sqft' had 785. Again, box plots showed the distribution (Fig. 11 and Fig. 12). The properties not resulting in a lead had higher means and many outliers. The mean for the 'Last Sold Price' was far greater in properties that did not result in a lead as well. This could imply that this company appeals stronger to property owners of smaller real estate.

The 'Tax Assessed Value' column required much effort and research to form. The raw datasets had a column ‘Total assessed value’ that had to be converted to its approximate tax assessed value. This was done using property tax laws in Florida, such as the Homestead Act and Save Our Homes (Moore, 2007)**.** The means of tax assessed values in properties that resulted in leads are relatively close, but properties that did not result in a lead had very high outliers, suggesting the company at hand did not appeal to properties with high tax assessed values. This was consistent with the idea that the company appealed best to smaller properties. Another possibility that could be derived from this is that a higher tax assessed value implied more recent home purchase. This is because tax assessed values rose when a property was purchased more recently, and intuitively a homeowner was far less likely to sell a new home.

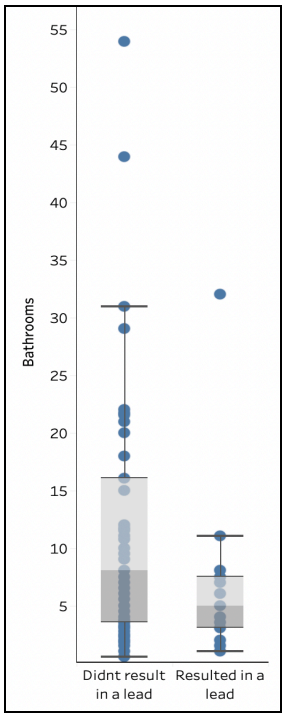

Fig. 9

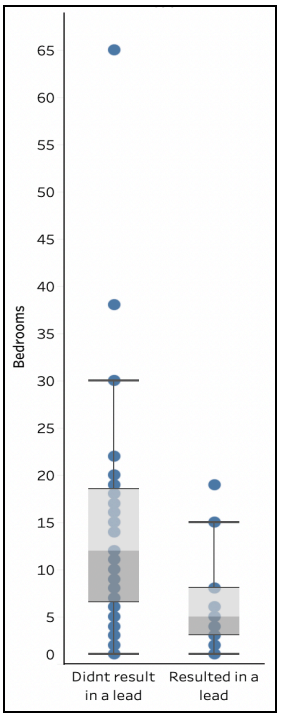

Fig. 10

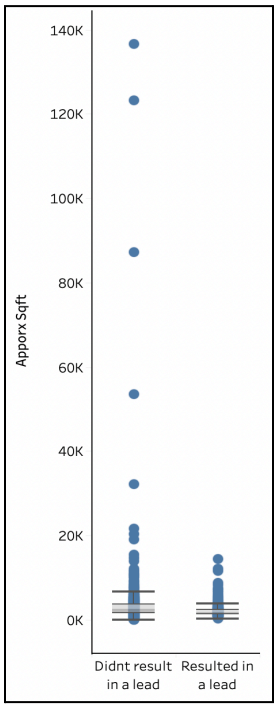

Fig. 11

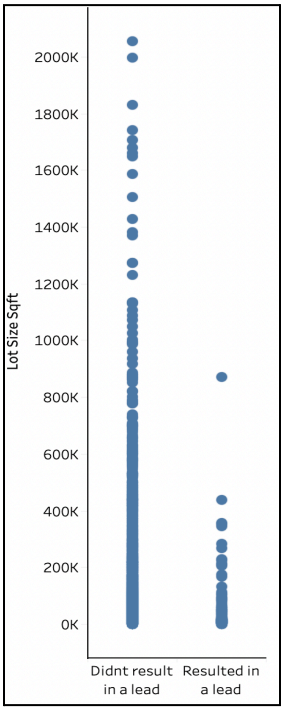

Fig. 12

The 'Effective Year Built' contained 651 null values. The line plot below showed the number of successful leads by the year the property was built (Fig. 13). The figure suggested properties built after the 1950's produced more leads.

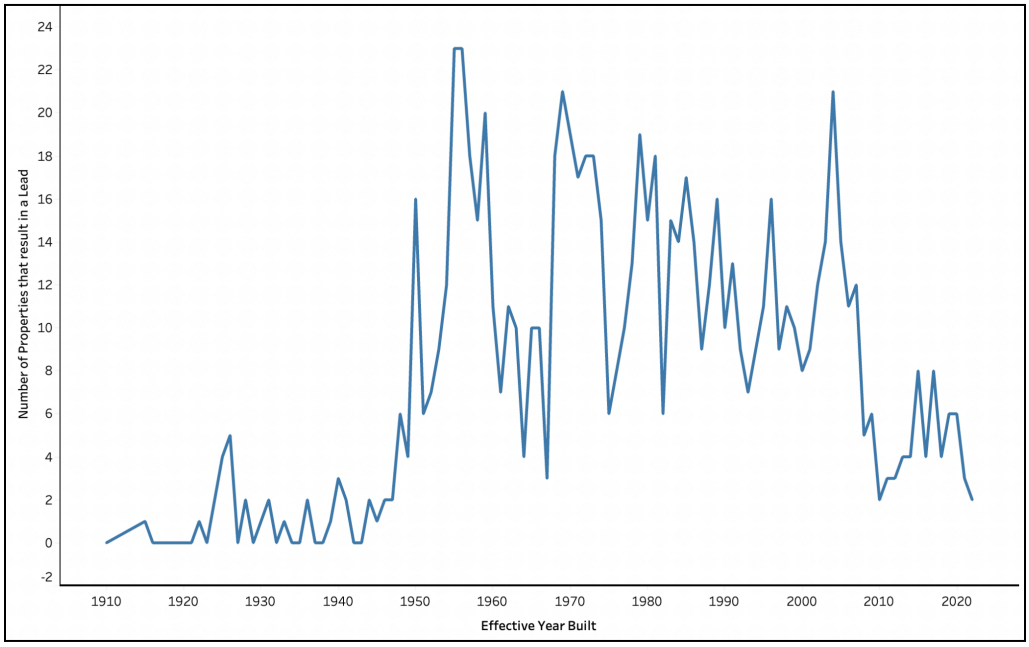

Fig. 13

In cases such as this one, where 99% of the data belonged to a single class, the model could achieve high accuracy simply by always predicting the majority class. To address this issue, one effective approach was to duplicate rows containing the rare target label, thereby balancing the dataset. This technique, known as oversampling, ensured that the model received sufficient training data for the minority class, improving its ability to predict all classes accurately and provided a realistic measure of its performance.

Results and Discussion

Table 1

Test Split .5	Model 1	Model 2	Model 3	Top hyperparameters with top 5 features
Test Accuracy:	0.9663617995897404	0.9790088992396317	0.9566451254684825	0.9561824267008036
Test Recall:	0.8741935483870967	0.8725806451612903	0.8612903225806452	0.6709677419354839
Test Precision:	0.20491493383742912	0.29676357652221613	0.1638539429272783	0.1362594169669178
F1 score:	0.3320061255742726	0.44289807613589854	0.27532869296210366	0.2265178328341955

As seen in table 1, Model 1 trained on 50% of the data did relatively ok. With a high accuracy of ~97% coupled with a recall of ~87%. Its low Precision and F1 score showed its low accuracy in false positives.

Model 2 outperformed the rest of the models with an accuracy of ~97% and a recall of ~87%. Although in the later digits the recall was lower than in Model 1, it was still a better model overall. The precision low at ~29% gave way to an F1 score of ~44%.

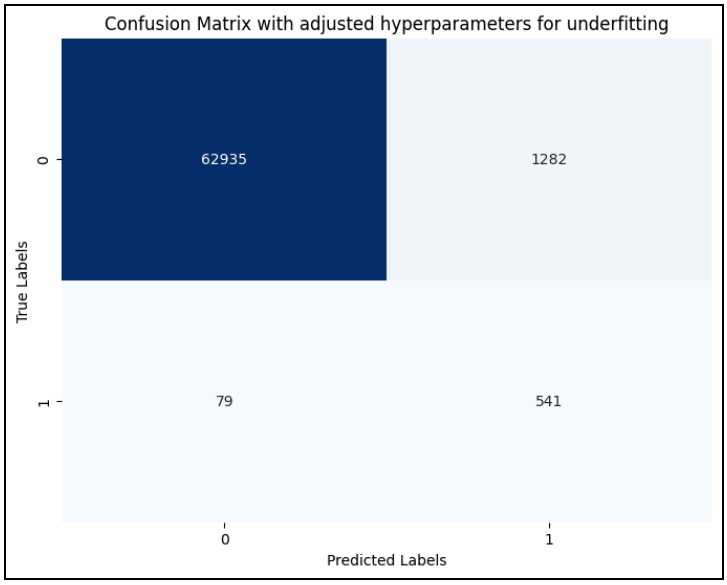

Fig. 14

Fig. 14 showed a confusion matrix summarizing this model's performance. Out of the 64,837 properties this model predicted, 63,476 were correctly labeled. 1,282 properties were incorrectly labeled a good lead, and 79 properties were falsely predicted to be a bad lead. Considering the purpose of these predictions, as mentioned before, the false positives were not nearly as costly as the false negatives.

Of the first three models, Model 3 did the worst with an accuracy of ~95%, recall of ~86%, extremely low precision of ~16%, and an F1 score at ~27%.

Fig. 15 showed the feature importance of Model 2, making these the most significant features when predicting leads. The top five features and their weights were 'Last sold price' -378, 'Effective year built' - 326, 'Tax Assessed Value' - 325, and 'Lot Size Sqft' -191, and 'Apporx Sqft' - 171.When a fourth model was created using the hyperparameters of Model 2, and only its top 5 features, its performance plummeted. The accuracy was ~95%, recall was ~67%, precision was ~13% and F1 the score was ~22%.

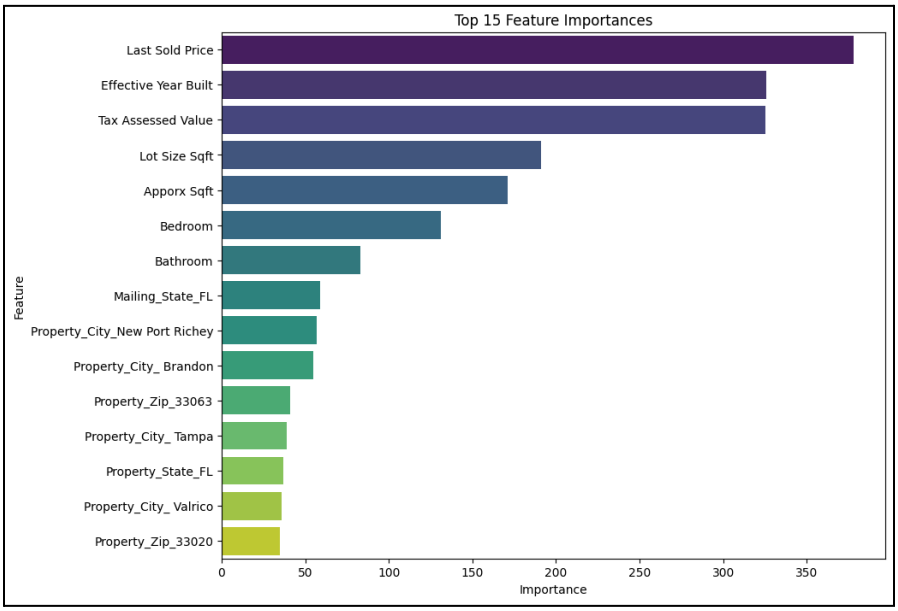

Fig. 15

Table 2

Test Split .2	Model 1	Model 2	Model 3	Top hyperparameters with top 5 features
Test Accuracy:	0.9718892817107873	0.9788747574908385	0.9620607889631386	0.9299417978012503
Test Recall:	0.8290322580645161	0.8870967741935484	0.7943548387096774	0.6733870967741935
Test Precision:	0.7002724795640327	0.7586206896551724	0.6118012422360248	0.40632603406326034
F1 score:	0.7592319054652882	0.8178438661710038	0.6912280701754385	0.5068285280728376

Table 2 represented the metrics of all models with a test split of .2. Model 1 with the test split .2 was a fine model with an accuracy of ~97% and a recall of ~82%. The precision was subpar at ~70%, causing the F1 score to be low as well at ~75%.

Just as the models made with 50% of the data, Model 2 did the best when trained on 80%, with an accuracy of ~98%, and the highest recall at ~89%. The precision was not excellent at ~75%, but as mentioned before it was not as important as the recall.

Fig. 16 showed a confusion matrix of the properties and their predictions. Out of 4,639 properties, 4,541 were predicted accurately. The 28 false negatives were the largest downside to the model, as each false negative represented a lost opportunity, whereas the 70 false positives would cost the company very little.

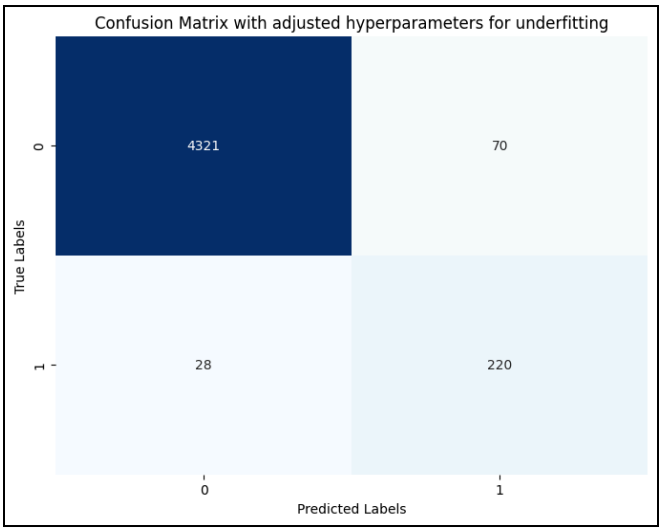

Fig. 16

Metrics on Model 3 proved it to be the worst of the first three models with the lowest accuracy, recall, precision, and of course F1 score.

The feature importance of Model 2 in this test split was similar to the one of the 50-50 test split. The top five features were the same, showing these were the most influential features consistently. The exact weights of the top five were not as great as the Model 2 with the .5 test split. ‘Last Sold Price’-255, ‘Tax Assessed Value’-246, ‘Lot Size Sqft’- 213, ‘Effective Year Built’- 198, and ‘Apporx Sqft’- 135.

When applied to the fourth model, using the same hyperparameters as Model 2 but only the top five features, its performance degraded exponentially. With an accuracy of ~92%, a recall of ~67%, precision of ~40%, and F1 score of ~50%.

The models trained on 80% of the data consistently did better than those trained on 50%. This performance disparity suggests that the models required a substantial amount of data to accurately capture and establish trends. By having access to more rows of data, the test split .2 models were better equipped to learn the underlying patterns and relationships within the dataset,

leading to more precise and reliable predictions. This finding highlighted the importance of larger training datasets to enhance the model's ability to generalize and perform well on unseen data.

As Model 2 was the strongest across both test splits, it became apparent that such predictions in properties needed a model that put much attention to small nuances in the data. Such hyperparameters in other cases may lead to overfitting, yet it was crucial in the predictions of the models.

The best model overall with the highest accuracy and recall was Model 2 trained on 80% of the data and tested on 20%. The combination of hyperparameters tuned for capturing small nuances and many rows of training data showed that the prediction of leads required attention to the slighter details. In both cases, when features were removed from Model 2 it worsened, proving testimony to the need for many features when predicting property leads.

Although the model was good, there is room for improvement. One area could be further parameter tuning with a gridsearch. A seemingly attainable goal would be to improve the precision, adjusting for false positives. This could be done by coupling this strong XGBoost model with another predictive model such as a logistic regression to filter the false positives. Using a k-fold would test if the model is overfitting in any iteration. Additional data could also be used to test the true accuracy of the model. Although it was clear that limiting it to 5 features was not beneficial, further analysis on feature importance can allow for subsequent filtering to ease computational power. There is much work to be done moving forward in this field, but this study has laid the groundwork to what could be a valuable asset in the world wholesaling real estate.

Conclusion

Throughout this study, significant progress was made across several key areas. Initially, comprehensive research was conducted on previous work in the field, providing a solid foundation for the project. The data was carefully analyzed and preprocessed. This ensured its suitability for model training, and gave light to how the company should view each column when trying to buy real estate. To address the imbalance in the target variable, oversampling techniques were employed, creating a more balanced dataset. Subsequently, multiple XGBoost models with varying hyperparameters and different test train proportions were developed to predict whether properties would produce leads.

Each model was evaluated to assess its performance, considering metrics of accuracy, recall, precision and F1 score, highlighting the strengths and potential areas for improvement in the models.

The best model overall was Model 2 trained on 80% of the data. The model used hyperparameters to create deep trees with many boosting rounds and a slow learning rate which lead to a model far more complex than the others. This model having the best performance taught that the data demands a meticulous and detailed approach. The model's ability to capture complex patterns and subtle variations indicated that a high level of precision and careful

consideration is necessary to fully understand and leverage the underlying data. Strong in accuracy and recall, the model is fit to do the job of narrowing the net cast by real estate wholesalers. However, adjusting to prevent false positives would greatly improve the model.

These efforts prove that it is possible to create a model for lead generation for a real estate wholesaling company, and represent the initial steps in developing an impactful resource in the field. By leveraging these findings, a real estate wholesaling company can significantly enhance its customer targeting strategies. Although further research is required to refine these models, the groundwork laid in this study is promising. This work can drive substantial improvements in the way the company identifies and pursues potential leads, ultimately transforming the approach to real estate wholesaling.[1]

[1] All relevant code can be found on Github: https://github.com/TamarSetton/Real-Estate-Wholesaling-Model

Annotated Bibliography

Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, O., & Afonso, C. (2018). Identifying Real Estate Opportunities Using Machine Learning. *MDPI.* https://www.mdpi.com/2076-3417/8/11/2321.

Case, K. E., & Shiller, R. J. (1988). The Behavior of Home Buyers in Boom and Post Boom Markets. *National Bureau of Economic Research.* https://www.nber.org/system/files/working_papers/w2748/w2748.pdf.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785.

Floyd, C. F., & Allen, M. T. (2002) *Real Estate Principles* (3rd ed., pp. 151-165, 350-358). Dearborn Real Estate Education.

Glower, M., Haurin, D. R., & Hendershott, P. H. (1997, November 13). Selling time and selling price: The impact of seller motivation. National Association of Home Builders.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from Class-Imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220-239. https://www.sciencedirect.com/science/article/abs/pii/S0957417416307175

Huang-Mei, H., chan, Y., Jia-Ying, X., Xue-Qing, C., & Zne-Jung, L. (2021) Data Analysis on the Influencing Factors of the Real Estate Price. *Universal Wiser Publisher.* https://ojs.wiserpub.com/index.php/AIE/article/view/966/608.

Krause, A., & Lipscomb, C. A. (2016) The Data Preparation Process in Real Estate: Guidance and Review. *Taylor & Francis Group. 19,* 15-42. https://www-jstor-org.library.saintpeters.edu/stable/24863180?searchText=real+estate+data&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dreal%2Bestate%2Bdata%26so%3Drel&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3Aefb5a48553e490669a9cb053541355b8&seq=1.

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence, 5*(4), 221-232. https://link.springer.com/content/pdf/10.1007/s13748-016-0094-0.pdf.

Moore, J. W. (2007). Examining property tax equity in Florida: Highlights from the doctoral dissertation proposal. Northcentral University.

Sarip, A.G., Hafez, M. B., & Daud, M. N. (2016) Application of Fuzzy Regression Model for Real Estate Price Prediction. *Malaysian Journal of Computer Science.* http://borneojournal.um.edu.my/index.php/MJCS/article/view/6889.

Sirgy, M. J. *Real Estate Marketing: Strategy, Personal Selling, Negotiation, Management, and Ethics* (pp. 3-6). Routledge.

Sknarev, D. & Trubnikova, N. (2020). The nature of the advertising image using the example of residential real estate advertising. *Revista Inclusiones*, 7, 550-566.

Xiuzhi, Z., Zhang, Y., & Zhijie, L. (2022). Online Advertising and Real Estate sales: evidence from the Housing Market. *Springer Link. https://link.springer.com/article/10.1007/s10660-022-09584-2.*

Zavadskas, E. K., Kaklauskas, A., & Banaitis, A. (2010) Real Estate's Knowledge and Device-based Decision Support System. *International Journal of Strategic Property Management. 3,* 271–282. https://www.ceeol.com/search/viewpdf?id=124910