Profitable Airbnb Investments: Where to Find Them? - Team 9

Diana Valencia, Jarrett Franklin, Joshua Kraus, and Warren Puy Arena

INTRODUCTION AND PROBLEM DEFINITION

Real estate investing is an ever evolving, highly competitive industry that encompasses the purchasing or rental of land and anything permanently attached to it, like a home [24]. As a market, it is impacted by a multitude of factors including demographics, the economy and government policies at a national and local level [25]. Plus, it has many players including property developers, individual investors, and investment firms and trusts among others [24].

In recent decades, the concept of short-term rentals (STR) has increased in popularity, opening new investment opportunities for players in the market [15]. STRs encompass a broad range of properties including vacation homes, furnished apartments, guesthouses, and rooms within a host's primary residence. The focus of these properties is on providing temporary accommodation for travelers or individuals in need of short stays, typically ranging from a few days to several weeks. With over six million active listings in 191 countries, Airbnb is the most well-known platform in the STR industry available in over 100,000 cities worldwide [12].

One of the initial steps when investing in income-producing real estate is to identify areas such as neighborhoods, zip codes, cities, counties, or states with above-average rental income potential or a favorable return on investment or yield. Multiple online free and commercial tools estimate individual property value and short-term rental (STR) income [2, 3, 4, 5, 6], yet none provide aggregate STR income potential information by area in one place or at no cost. Given this, the objective of this project was to develop a model that can identify and predict geographical areas with favorable STR investment returns in the United States.

Literature Review

The main factor determining rental profitability is the yield, defined as the received rental fees divided by the rental property value. Prediction of individual real estate property prices and rent has been studied for decades. According to Park and Bae (2015), the two main research trends in modeling house prices are regression-based approaches and machine learning (ML) techniques. Their paper included a historical account of methods used since 1996 [16]. Baldominos et al. (2018) built housing price prediction models based on ensembles of regression trees, k-nearest neighbors, support vector machines for regression, and multi-layer perceptrons, which highlighted the outperformance of ensembles of regression trees [17]. Similarly, Mrsic et al. (2020) developed apartment price prediction models based on random forest, gradient boosting, AdaBoost, decision trees, k-nearest neighbors, linear regression, and XGBoost algorithms, noting the outperformance of the XGBoost algorithm [18]. All these models used descriptive predictors such as square footage, number of bedrooms and bathrooms, and public-school ratings as independent variables. Additionally, the findings presented provided a foundation of the explanatory factors and ML techniques that could predict real estate prices and rent.

Other researchers demonstrated the importance of market imbalances and socioeconomic variables when predicting real estate prices and rent. According to McCue and Belsky (2007), income and employment growth and the user’s cost of capital are the primary drivers for long-run real estate price appreciation. In the short-term, however, deviations from the long-term equilibrium can occur when changes to the supply and demand of housing are introduced [19]. In their work, Zhu and Sobolevsky (2018) commented that house prices are associated with local socioeconomic factors, such as income level, unemployment rate, education level, school quality, and crime level. They introduced the use of local socioeconomic factors extracted from real-time digital census datasets, such as New York city 311 complaint calls, crime complaints, and taxi trips data, and provided evidence that their use could improve the performance of house price models [20].

Finally, work from Boeing and Waddell (2017), Grimes and Aitken (2007), and Puy Arena et al. (2022) on long-term-rentals (LTR) and work from Adamiak et al. (2019) and Sarkar et al. (2017) on STR markets also highlighted the presence of different rental profitability based on regional and socioeconomic factors [21, 22, 1, 13, 14]. This fact presented an opportunity for creating a model to predict and identify geographical areas with above-average STR income potential.

Research Questions

The primary and secondary research questions for the project were:

1. What are the key factors that drive STR profitability at the geographical area level?
2. Is it possible to accurately predict STR profitability of an area using analytical tools?

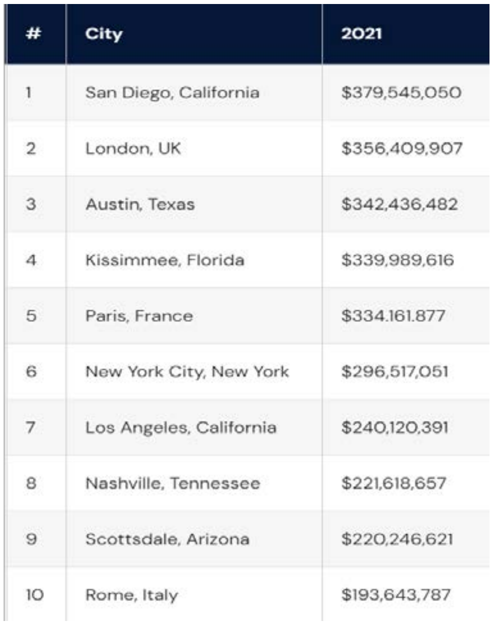
Although not the primary goal, we also sought to answer the following supporting research questions:

1. How do key factors that drive STR returns compare with those that influence LTR profitability?
2. What role do the population socio-economic factors play in determining STR profitability? What is the socio-economic profile of the areas with above-average STR investment returns?

Business Justification

By exploring statistical data about the STR market and Airbnb in particular, we strengthened our belief that there is a significant business opportunity in this project. As can be observed in the table to the right (Figure 1), according to Search Logistics, 7 out the top 10 cities in the world with the highest Airbnb gross revenue as of 2021 were in the U.S. Additionally, Search Logistics states that "the U.S. also generated more revenue than the next 9 countries combined" [12]. Given that our goal was to provide investors with tools to maximize their profits, exploring the US Airbnb market appears to be a worthwhile venture.

If successful, real estate investors and firms such as private equity funds, asset management corporations, and investment trusts could rely on this tool to identify areas with favorable investment returns. Using this tool could result in a more efficient allocation of their investment capital in a concentrated geographical area and overall, higher investment returns.



#	City	2021
1	San Diego, California	$379,545,050
2	London, UK	$356,409,907
3	Austin, Texas	$342,436,482
4	Kissimmee, Florida	$339,989,616
5	Paris, France	$334.161.877
6	New York City, New York	$296,517,051
7	Los Angeles, California	$240,120,391
8	Nashville, Tennessee	$221,618,657
9	Scottsdale, Arizona	$220,246,621
10	Rome, Italy	$193,643,787

Figure 1: Airbnb gross revenue by city from Search Logistics [12].

METHODOLOGY

Datasets

Airbnb Data

A considerable amount of time and effort was invested to gather Airbnb regional income and supply and demand data. One of the main challenges was that Airbnb does not release any data to the public, and the readily available data from different individuals and organizations are missing factors which we hypothesize will be key to our model. Considered alternatives to our approach included:

- Utilizing U.S. Airbnb Open Data [8] published in Kaggle by Kritik Seth, however, this data source was limited to a handful of urban areas in the U.S. Additionally, this data set did not provide occupancy rate data for geographical regions which would have required heavy manipulation to generate usable features including but not limited to mapping listing locations to zip codes.

- The team reached out to Mashvisor [3], Rabbu [4], Airbitics [5], and AirDNA [6], four online STR data providers requesting access to their Airbnb data for use in our project. They all display on their website many of the attributes the team hypothesizes will be key to our model grouped by zip code, which was the desired geographical area unit to analyze. Unfortunately, all data providers quoted an unattainable fee, ranging from $1,000 to $8,000 to extract and provide the requested data.

Given the absence of a complete, low cost, and readily available Airbnb dataset, the team opted to scrape the necessary data from Rabbu's website. Rabbu was selected as the source because out of all Airbnb data providers, Rabbu was the only one that displayed all relevant attribute values on their website (see Figure 9 in the appendix). To make comparison with the LTR market easier, we decided to pull Airbnb data for all zip codes contained in the Zillow Observed Rent Index (ZORI) [7]. There are 6,859 zip codes in ZORI which constitute a good representative sample of the 33,120 populated U.S. zip codes (20.71%).

Data scrapping was done in Python using Selenium. We developed an automation script that entered each target zip code on Rabbu's web page [4], deselected all property types but "Entire Homes", extracted the relevant values from the page code, and saved the extraction to a Python dictionary. Lastly, our code created a Pandas data frame from the dictionary and wrote the results to a .csv file. We were able to find and extract key attribute values for 4,384 out the 6,859 targeted zip codes.

U.S. socio-economic data

For the zip code socio-economic variables, we used the raw data extracted from the U.S. Census American Community Survey (ACS) [10] by Puy Arena et al. (2022) [1]. The dataset provided included data for the years 2014 to 2020 pulled from ACS data profiles DP02, DP03, DP04, and DP05. The dataset has socio-economic attributes for 33,120 zip codes (see Figure 12 in the appendix).

Zillow Home Value Index (ZHVI)

This dataset was directly downloaded from Zillow [7]. It was used as a proxy for the zip code "typical" home value or median value for the middle tier, smoothed and seasoned adjusted cut of all homes in the zip code. It contains a monthly time series of real estate property values from January 2000 to May 2023 (see Figure 10 in the appendix).

Zip Code Database

This dataset was directly downloaded from the provider's website [9]. Their data sources include the USPS, U.S. Census Bureau, and the Internal Revenue Service (IRS) among others. It was used to extract zip code location information such as the longitude, latitude, county, city, and state (see Figure 11 in the appendix).

Key Variables

Independent variables: Real estate property metadata including location and average property value, occupancy rate, daily rate, and average number of active listings. Socio-economic variables including income level, unemployment rate, education level, minority percentage, percentage of the population with income below the poverty level, number of vacant rental units, and number of occupied rental units by renters.

Dependent variable: STR profitability by zip code. A zip code was labeled as profitable when its annualized rental income yield, its annualized Airbnb net income divided by the average property value in the zip code, is 18% or higher. This profitability threshold is based on the empirical real estate "1% rule" where investors aim to earn at least 1% of the property purchase price per month to cover carrying costs and generate a reasonable return on investment [11]. An additional 3% paid to Airbnb for booking fees and another 3% to cover other STR costs such as local transient occupancy tax, property management fees, higher maintenance cost, utilities, and supplies was also added to this threshold.

The hypothesis was that location and STR property supply and demand variables, such as property value, rental prices, average number of listings and average occupancy, would be the most influential factors in determining the region's STR profitability.

Except for location, it was expected that our analysis would show that the most relevant factors driving a region's STR profitability differ from those driving its LTR profitability, and that this difference may arise due to distinct supply and demand dynamics in both markets.

Data Exploration and Preparation

The goal was to create a model that helps investors with the initial screening of areas for investment. Consequently, the focus was on providing a binary response that classifies a zip code between profitable, defined as a zip code yield equal or greater than 18%, or not. To accomplish this, as described in the previous section, profitability labels were assigned to the training dataset.

Our preliminary exploratory data analysis of the training dataset showed that U.S. zip code yield data follows a right-skewed distribution, with most observations falling between 3 and 12 %. Profitable zip codes (yield ≥ 18%) were only 5.45% of all zip codes (minority class). Figure 2 below shows the zip code yield data distribution. Figure 3 below shows a plot of the correlation matrix, or heatmap, for the dependent variable, or the yield, and key supply and demand independent variables. These results show that there are enough profitable data points to train a model and that other independent variables such as location and socio-economic factors should be considered to predict zip code profitability.

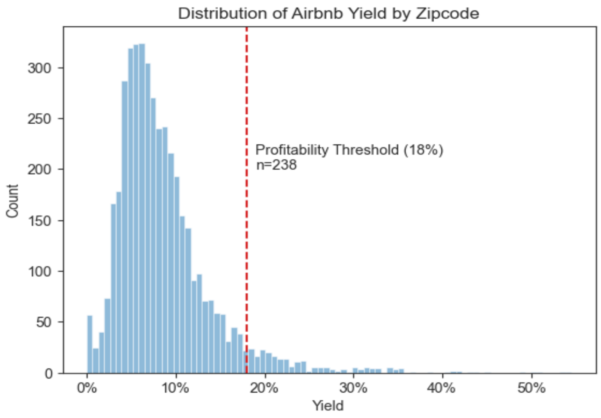


Figure 2: Zip code gross yield data distribution.

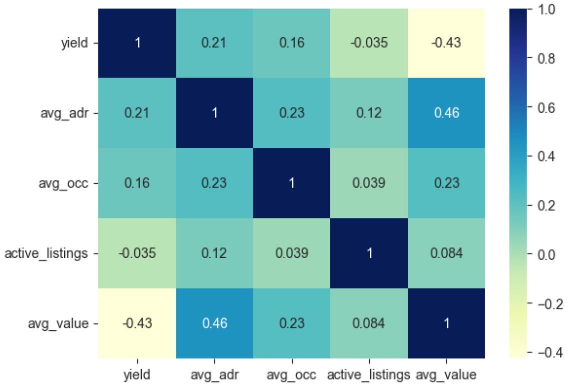


Figure 3: Heatmap for dependent and independent variables.

We first joined the U.S. socioeconomic dataset with the zip code database to add zip code's type and geographical information to the socioeconomic data. We removed zip codes starting with 006 and 009 due to missing data (131zip codes). After that, we extracted location information as well as the last 12 months of average home prices from the ZHVI dataset and created a new column with the average home value for the last 12 months. The resulting data was then combined with a "clean" version of the scrapped Airbnb data. At this point, we had the two components needed to calculate the zip code yield (Airbnb's seasonalized annual projected revenue and Zillow's average home value) in the same dataset. After creating a new column to store the yield value, we performed the preliminary exploratory analysis described in the previous paragraph. Finally, we combined the socioeconomic and the Airbnb yield datasets. After removing rows with missing yield data or yield = 0 (60 rows), we ended up with 4,302 zip code records with 7 years of socioeconomic predictors and current Airbnb supply and demand variables (including the Airbnb yield) for a total of 156 variables.

We then proceeded to remove zip codes with missing data and a population below 1,000. This population threshold was implemented to ensure a sufficient presence of both supply and demand within the area, enabling market forces to effectively regulate the housing market equilibrium. Also, a new column was created to label the zip code as profitable if its yield was 18% or greater. We only selected socioeconomic factors for the last three available years (2018 to 2020). This decision was made based on the results of previous work done by Gilling et al. (2021) and Puy Arena et al. (2022) on the "lagged" effect that geographical socioeconomic factors changes have in the area real estate pricing and income. They found that there is a three-to-four-year lag between the changes and the effect in real estate income [23, 1]. This resulted in a dataset with 4,295 zip codes and 73 variables, 236 of them being "profitable" (5.49%).

Before feeding the data into the machine learning pipeline, we carved out a small random "unseen" dataset from the labeled data (10% of the data, 430 zip codes, 21 being from the minority class). Finally, to deal with the skewed distribution of the zip code profitability label in the training dataset (imbalanced classes), we created three additional versions of the training dataset by resampling the minority class data. We used the following "over-sampling" algorithms: Random Over-Sampling (ROS), Synthetic Minority Oversampling Technique (SMOTE), and Generative Adversarial Networks (GANs) synthetic data generation [20, 28]. Data cleaning, integration, and preparation were performed with Python, utilizing PySpark, Numpy, Pandas, sklearn, imblearn, and tabgan libraries.

Feature Selection

Besides removing all socioeconomic factors for the years 2014 to 2017, we decided to program our classification pipeline to ignore the seasonalized Airbnb revenue attributes and number of active hosts. The former since the annual version was directly used to compute the zip code yield and the fact that we already had the average daily rate in the model as a proxy for STR income. We removed the number of active hosts because we already had the number of active listings as a proxy for STR supply. These two variables are highly correlated. Following along these lines, we also programmed our classification pipeline to remove highly correlated factors using a correlation threshold of 0.9. The reason was to address multicollinearity and help reduce the complexity of the model without impacting its predictive power.

Models

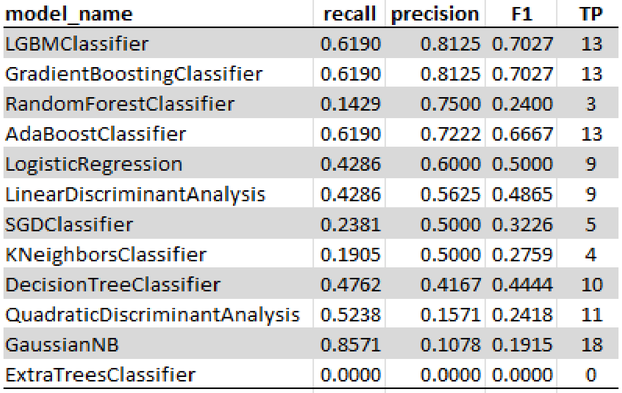
With the assignment of profitability labels to the training dataset ("True" for zip code yield $\geq$ 18%, "False" otherwise), predicting a zip code STR profitability based on zip code STR supply and demand and socioeconomic data becomes a machine learning classification problem. We addressed this problem by creating a "classification" pipeline using the PyCaret Python library to select and tune the "best" model. Since profitable zip codes were the minority class, we also used resampled versions of the training dataset as described in the "Data Exploration and Preparation" section above. We trained and compared results for the following "classifier" algorithms included with PyCaret's classification module: Logistic Regression (LR), K Nearest Neighbors (KNN), Naives Bayes (NB), Decision Tree (DT), Random Forest (RF), Extra Trees Classifier (ET), Support Vector Machine Linear Kernel (SVM), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Ada Boost (ADA), Gradient Boosting (GB), and Light Gradient Boosting (LGB). Based on the results of previous work done by Gilling et al. (2021) and Puy Arena et al. (2022), we decided to predict zip code profitability for the current year, 2023 (three-year lag), after training and comparing the previously mentioned classifiers with the last three years of available socioeconomic feature data (2018 to 2020) [23, 1].

Our initial or "base" model configuration was as follows:

- Training data size: The data was split 70% or 2705 zip codes for training, and 30% or 1160 zip codes for model validation. This split was adopted to provide a good tradeoff between reducing bias by using a substantial portion of data to train the model, while still retaining an adequate sample size to accurately estimate and compare performance metrics for each model tested. This split did not include the test data set mentioned previously, which was withheld for further examination of the best model only.

- Normalization: To improve performance and stability of some of the algorithms used, we set this variable to “True”.
- Fold: Given the small number of the minority class data points, we decided to decrease the number of cross-validation folds from 10 to 5.
- Transformation: Initially set to “False”, this was one of the parameters changed during model fine-tuning to see if we could improve model performance.
- Remove multicollinearity: As mentioned in the “Feature Selection” section above, this setting was setup to “True” with a correlation factor threshold of 0.9.
- Feature selection: Initially set to “False”, this was one of the parameters changed during model fine-tuning to see if we could improve model performance.
- Polynomial features: Initially set to “False”, this was one of the parameters changed during model fine-tuning to see if we could improve model performance.
- PCA (dimensionality reduction): Initially set to “False”, this was one of the parameters changed during model fine-tuning to see if we could improve model performance.

Given that misclassifying profitable areas could result in financial losses for investors, for feature selection and model performance comparison, we used precision for the minority class as our primary model comparison metric. For each combination of model parameters and training data tested, we selected the model with the best combination of test (unseen data) precision for the minority class, recall and number of true positives (TP). Figure 13 below shows a summary of model performance metrics after running our classification pipeline with the “base” model configuration. Figure 14 shows the classification performance of the “best classifier” (LGBM Classifier) for the base configuration. There are 21 profitable zip codes in the test (unseen) dataset.



model_name	recall	precision	F1	TP
LGBMClassifier	0.6190	0.8125	0.7027	13
GradientBoostingClassifier	0.6190	0.8125	0.7027	13
RandomForestClassifier	0.1429	0.7500	0.2400	3
AdaBoostClassifier	0.6190	0.7222	0.6667	13
LogisticRegression	0.4286	0.6000	0.5000	9
LinearDiscriminantAnalysis	0.4286	0.5625	0.4865	9
SGDClassifier	0.2381	0.5000	0.3226	5
KNeighborsClassifier	0.1905	0.5000	0.2759	4
DecisionTreeClassifier	0.4762	0.4167	0.4444	10
QuadraticDiscriminantAnalysis	0.5238	0.1571	0.2418	11
GaussianNB	0.8571	0.1078	0.1915	18
ExtraTreesClassifier	0.0000	0.0000	0.0000	0

Figure 13: Model performance metric summary (base configuration, data = test (unseen dataset)) sorted by precision.

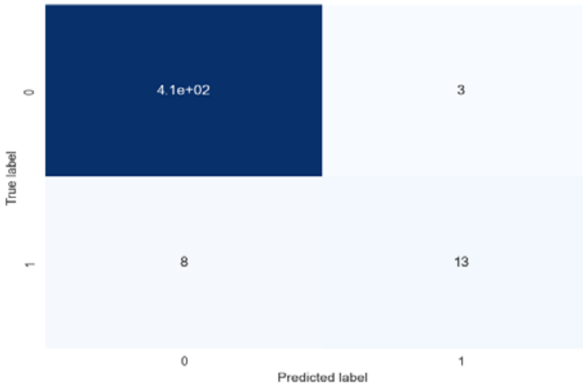


Figure 14: Confusion matrix for LGBM Classifier model (base configuration, data = test (unseen dataset)).

Interestingly, the three out the four top identified models were from the category of boosting algorithms: Light Gradient Boosting Machine, Gradient Boosting Classifier, and Ada Boost Classifier. In boosting algorithms, a random sample is selected, fitted to a model, and trained sequentially. Boosting algorithms improve the prediction power of weaker classification models by each model trying to compensate for the weaknesses of its predecessor. Given this, with each iteration the weak rules from each individual classifier are combined to form one, strong prediction rule [27]. This behavior of iteratively correcting mistakes of prior training cycles can yield boosting algorithms capable of capturing complex relationships like those in the STR market, therefore explaining their strong performance.

The best model for the base case was a Light Gradient Boosting Machine (LightGBM) model, which can be described as a distributed, efficient gradient boosting framework utilizing tree-based learning which is designed to handle large-scale data of high dimensionality and identify nonlinear relationships between predictor and the dependent variable [26]. Although, this project's initial dataset was comprised of only 4,295 records and 73 dimensions, we suspect complex relationships are present between STR profits and the selected features.

EXPERIMENTS AND EVALUATION

Starting with the base model described in the previous section, multiple attempts were made to improve the selected model performance metrics. Specifically, eleven iterations of the classification pipeline were conducted in an attempt to improve performance. Tuning techniques involved altering the training size from 0.7 to 0.8, implementing outlier removal, data transformations, feature selection, polynomial features, and dimensionality reduction through Principal Component Analysis (PCA), as well as applying resampling techniques previously mentioned. The synthetic data generation techniques attempted were Random Oversampling (ROS), Synthetic Minority Over-sampling Technique (SMOTE), and Generative Adversarial Networks (GANs). Among the tested modifications, only an increase in the training data size from 70% to 80% resulted in improved performance. Figures 15 and 16 in the appendix summarize these results.

Based on these results, the two top performing models were chosen to continue with analysis. These models were based on the Ada Boost Classifier and the LightGBM, both trained with 80% of the training dataset. To enable a fair comparison, both models were finalized by training them with 100% of the training data (no train/test split). Subsequently, their performance was evaluated on the hold-out test dataset. Both models correctly identified an equal number of profitable zip codes, however, the Ada Boost Classifier exhibited superior precision for the minority class. To further validate these findings and mitigate concerns about overfitting, each of the two selected models were fitted and tested ten times, each time predicting a different random unseen sample. The analysis revealed that both models demonstrated generalization capabilities, as their precision standard deviation was below 10%. The average precision for the minority class over the ten runs was 0.8350 for the Ada Boost classifier and 0.8110 for the LightGBM classifier. Figures 17 and 18 in the appendix summarize these results.

Finally, to make an informed decision about the most suitable model, both trained models were tested on all available data, including the hold-out dataset. Surprisingly, the LightGBM model correctly identified 228 profitable zip codes, whereas the Ada Boost Classifier only managed to identify 185. Based on this result, we selected the model based on the LightGBM as our "best" model.

Observations

1. The LightGBM in conjunction with a training data size of 80%, predicted the minority class (profitable zip codes) with an average precision of 0.8110 and produced the highest number of true positives (228) when predicting the entire dataset when compared to the Ada Boost classifier (185).
2. The results from the LightGBM model revealed that the most important predictors were average daily rate, average home value, and average occupancy (see Figure 4 below). While other variables also had importance in the classification model, the difference between importance from variable to variable decreased significantly after the first three predictors. This result was expected as these variables have a strong influence on the zip code average income and property value and confirmed our initial hypothesis that they were key factors in identifying profitable zip codes. Based on this result, we decided to further examine the effect of these three variables on zip code profitability in a separate section below.
3. The zip code geographical location played a significant role in determining profitability, as evident from the top ten feature importance plot (figure 4) which included both zip code latitude and

longitude. Figure 5 shows profitable zip code locations on a U.S. map. Notably, the majority of profitable zip codes are concentrated in the eastern half of the country, with only 23 (9.75%) located in the West and Southwest regions. This finding aligns with a study on LTR profitability by Puy Arena et al. (2022). They reported a similar geographical pattern with only 4.45% of profitable zip codes found in the West and Southwest regions [1].

4. In the top ten feature importance plot (figure 4), three of the four crucial socioeconomic factors were associated with real estate property supply, including rental vacancy rates and total vacant units. This outcome was as anticipated since a surplus of supply in the STR market often leads to a decrease in average daily rates and occupancy, ultimately reducing rental income and expected return or yield. Interestingly, a study on the LTR market [1] revealed a different trend, with four out of the top five socioeconomic factors influencing returns linked to population income level variables. These variables included population below the poverty line, minority population, median household income, and percentage of the population with a bachelor's degree.

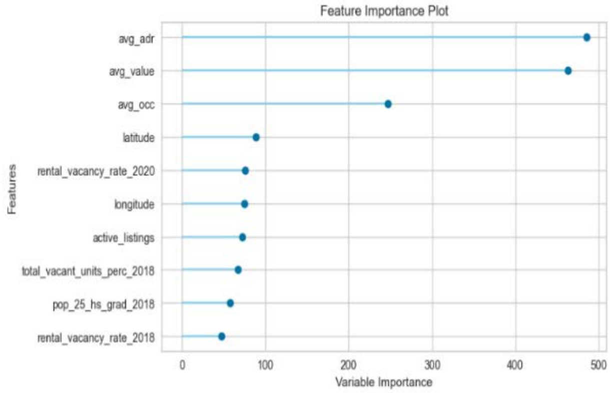


Figure 4: Feature importance plot for selected model.

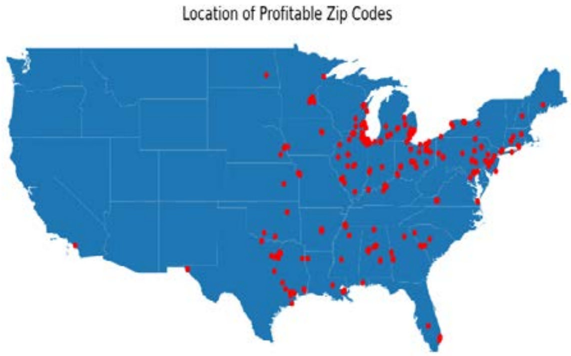


Figure 5: Location of profitable zip codes.

Key Model Features Analysis

The most important predictor in the LightGBM model was the average daily rate, defined as the average rate posted on Airbnb by zip code for a specified period [4]. To understand the effect of average daily rate when identifying profitable zip codes for short-term rentals, the distribution of average daily rate was examined. To facilitate comparison, violin plots were created for each profitability class assigned by the selected model (0 = “Not Profitable”, 1 = “Profitable”). As it can be seen in Figure 6, the distributions of average daily rate for both classes were right-skewed, with long right tails. This revealed that for both profitable and non-profitable zip codes average Airbnb prices tend to be skewed towards lower prices, with some more expensive options present in both classes. Most notably, the median average daily rate for profitable zip codes was higher than that of non-profitable ones, with a rate of $234.50 compared to $186 per night. This suggested that, up to a certain extent, charging more per night for short-term rentals may be related to higher yield. Further examination of this relationship would be required to evaluate statistical significance and possibly identify thresholds in increasing daily rates to help investors maximize returns.

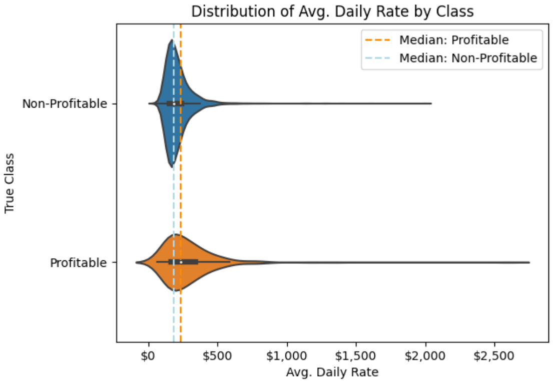


Figure 6: Average daily rate by class.

The second most important predictor in the model was average home value, defined as the average value of the property retrieved from Zillow. Similar to average daily rate, as it can be seen in Figure 7 the distribution of average home value for both classes was right skewed, suggesting that home values tend to skew cheaper for all zip codes. Different from the previous examinations, however, only the non-profitable class had a long tail for more expensive homes. Even though it is important to highlight there was low exposure of these more expensive properties in the data, making it difficult to draw concrete conclusions, this suggested that high value properties may not be the most profitable for short-term rentals. This finding is reinforced by the fact that the median average home price for non-profitable zip codes was higher than that of profitable ones, at an average price of $398,520 for non-profitable zip codes compared to $164,673. This finding can be explained by the fact that the higher the value of a home, the higher the average daily rate charged on Airbnb needs to be to realize a return on investment. Given that it has already be determined that Airbnb prices tend to be skewed cheaper for both classes, increasing the average daily rate to account for higher home value may make it more difficult to realize returns as customers may push for cheaper Airbnb rates.

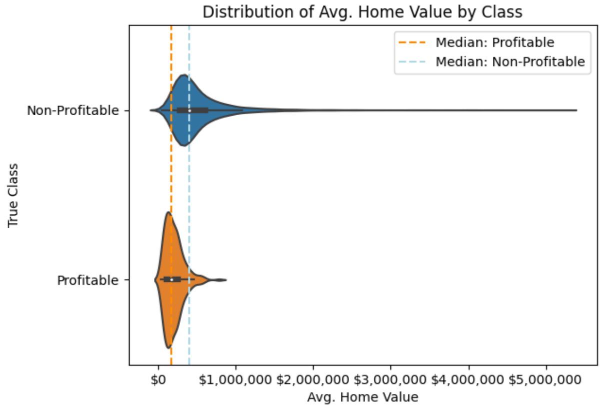


Figure 7: Average home value by class.

The final variable examined was average occupancy, defined as the average percentage of time Airbnb's are booked for each zip code, which was the third most important variable in the selected model. As it can be seen in Figure 8, the distribution of average occupancy for both classes was fairly normal, though slightly short-tailed. Notably, the average occupancy of profitable zip codes is slightly higher than that of non-profitable ones, at 57% compared to 53%. This is to be expected, however, as higher occupancy rates lead to more return for investors, making it easier to reach the profitability threshold of 18%. While the relationship between Airbnb features and average occupancy rates were not examined in this study, further exploration of these factors could help investors better understand which features could help them realize higher occupancy rates, therefore leading to higher yield.

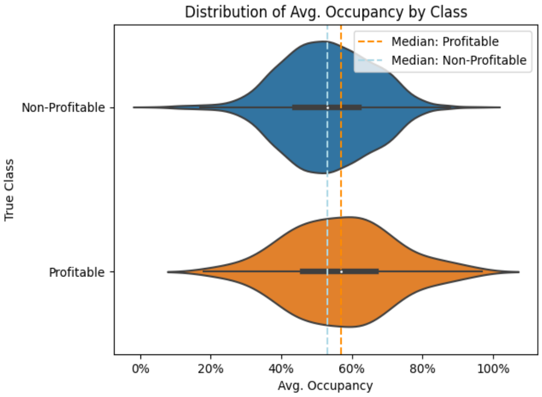


Figure 8: Average occupancy by class.

CONCLUSIONS AND DISCUSSION

In conclusion, the LightGBM model trained in this project successfully predicted profitable zip codes as observed in the evaluation section. The model achieved a precision of 81.1% which could allow investors to identify zip codes that can serve as a starting point in their STR property search with a reasonable degree of confidence. The primary focus of our model was to achieve a high level of precision when predicting profitable zip codes, even if it translated to reduced accuracy. This strategy led to a considerable percentage, around 40%, of profitable zip codes to be misclassified as unprofitable. Although this outcome represented a "loss" in terms of missed profitable opportunities, it was a deliberate trade-off to optimize the precision of predictions for the profitable class. This decision was made to safeguard against mistakenly classifying an unprofitable area, as such misclassification could potentially lead to significant financial losses for investors.

Although this is a predicting model and the intention of this project is not to conclude causality, we found there is a strong relationship between STR yield and the top three contributing factors: average daily rate, average home value, and average occupancy. As described in the observations section, it was found that the most profitable zip codes possess a median home value lower than that of nonprofitable zones, but a higher daily rate and occupancy levels. Armed with this knowledge, even if investors had no access to the list of profitable zip codes resulting from this project, they could set out on their search for regions with on average inexpensive properties where they can charge a competitive rate. Additionally, they could also further explore what the drivers of occupancy levels are to maximize their profits.

In comparison to LTR profitable regions prediction, both LTR and STR profitability models appear to point to the eastern part of the country as having the most potential for real estate investment. Given these comparable results, further examination is required to determine why eastern regions may be more profitable for rental real estate investments. When considering the key features identified in our model, it is possible that property value is lower in the East and Midwest regions compared to that of the West, making it easier to realize a return on investment. It is also possible that other features could be driving a region's profitability which could explain this difference, such as factors related to the region's desirability for travel and population density or concentration. In contrast, the two models differ when it comes to the most relevant socioeconomic factors driving a region's profitability, where STR is related to the real state supply in the market and LTR is related to population's income levels.

Future Work

For this tool to stay relevant to investors, the data needs to remain accurate and current. Consequently, future work should include periodic updates of the data sources. Ideally, this process would be automated using APIs to retrieve data on a schedule. Because the sources used included data owned by private entities Zillow and Rabbu, future work should include building a partnership with these or similar providers to utilize their data commercially.

Additionally, improving the tool's usability is crucial to attract and retain investors. Creating a user-friendly visualization tool with a map and intuitive filters will help investors identify patterns and profitable zones, empowering them to make well-informed decisions.

Finally, this project did not account for various other influential factors that could significantly impact the STR market. These factors include, but are not limited to:

- Macroeconomic: Nationwide economic changes such as GDP growth, inflation rates, and interest rates which could impact the overall profitability potential of the STR market.
- Policies and Regulations: Government policies at a local, state, and national level such as tax laws, zoning laws, STR moratoriums which could impact the feasibility of locating short-term rentals in specific geographic regions.
- Local Market: Characteristics such as accessibility, convenience, and safety could make a region more attractive for STR, increasing demand and potential earnings. On the other hand, these factors could also drive increases in property value, cutting into an investor's potential yield.
- Hotel Supply and Pricing: A supply of competitively priced hotels serves as a direct competitor for STR services such as Airbnb, and related factors could impact the decisions of short-term rental seekers.
- Other STR Data: Though Airbnb is the largest STR platform, factors related to other competitors such as VRBO could increase robustness of the training data set [12].

These factors were excluded from the project for several reasons including data availability, time constraints, and practicality. The goal of the team was to achieve high precision with a simple model that would add significant value to investors, however, it is recommended that the predictors listed above be evaluated in greater depth as part of future work.

REFERENCES

[1] Puy Arena, W., Natta, D., McLaughlin, P., & Kafity, G. (2022). Rental Real Estate Opportunity Locator.

[2] *Real estate, apartments, Mortgages & Home values. Zillow. (n.d.). https://www.zillow.com/*

[3] *Your search for investment property begins and ends here. Mashvisor. (n.d.). https://www.mashvisor.com/*

[4] Free Airbnb Calculator & Data Analysis - instant projections. (n.d.-a). https://data.rabbu.com/

[5] *The smartest way to optimize your short-term rentals*. Airbtics. (2023, June 15). https://airbtics.com/

[6] *Your short-term rental. Our data. More revenue.* AirDNA. (n.d.). https://www.airdna.co/

[7] *Housing Data*. Zillow. (2023, April 25). https://www.zillow.com/research/data/

[8] Seth, K. (2023, April 14). *U.S. Airbnb Open Data*. Kaggle. https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data

[9] Zip code database - ZIP code list (updated for 2023). (n.d.-b). https://www.unitedstateszipcodes.org/zip-code-database/

[10] Bureau, U. S. C. (n.d.). Explore census data. https://data.census.gov/table?d=ACS+5-Year+Estimates+Data+Profiles

[11] *Breaking down the 1% rule in real estate.* Breaking Down The 1% Rule in Real Estate | Rocket Mortgage. (n.d.). https://www.rocketmortgage.com/learn/1-rule-real-estate

[12] *Airbnb statistics [2023]: User & market growth data*. SearchLogistics. (2023, June 12). https://www.searchlogistics.com/learn/statistics/airbnb-statistics/#:~:text=Airbnb%20is%20a%20big%20platform,have%20Airbnb%20listings%20in%20them

[13] Adamiak, C., Szyda, B., Dubownik, A., & Garcia-Alvarez, D. (2019). Airbnb Offer in Spain-Spatial Analysis of the Pattern and Determinants of Its Distribution. ISPRS International Journal of Geo-Information, vol. 8, no. 3, 2019, p. 155.

[14] Sarkar, A., Koohikamali, M., & Pick, J. (2017). Spatiotemporal Patterns and Socioeconomic Dimensions of Shared Accommodations: The Case of Airbnb in Los Angeles, California." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. 4, no. 4, 2018, pp. 107–14.

[15] Property Management Vacation Rental Hosting, et al. "*The History of Short-Term Rentals."* Keycafe Blog, 27 June 2023, blog.keycafe.com/the-history-of-short-term-rentals/.

[16] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert systems with applications, 42(6), 2928-2934.

[17] Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernárdez, Ó., & Afonso, C. (2018). Identifying real estate opportunities using machine learning. Applied sciences, 8(11), 2321.

[18] Mrsic, L., Jerkovic, H., & Balkovic, M. (2020, March). Real estate market price prediction framework based on public data sources with case study from Croatia. In Asian Conference on Intelligent Information and Database Systems (pp. 13-24). Springer, Singapore.

[19] McCue, D., & Belsky, E. S. (2007). Why Do House Prices Fall?: Perspectives on the Historical Drivers of Large Nominal House Price Declines (No. 3). Joint Center for Housing Studies, Graduate School of Design [and] John F. Kennedy School of Government, Harvard University.

[20] Zhu, E., & Sobolevsky, S. (2018). House price modeling with digital census. arXiv preprint arXiv:1809.03834.

[21] Boeing, G., & Waddell, P. (2017). New insights into rental housing markets across the United States: Web scraping and analyzing craigslist rental listings. Journal of Planning Education and Research, 37(4), 457-476.

[22] Grimes, A., & Aitken, A. (2007). House Prices and Rents: socio-economic impacts and prospects.

[23] Gilling, G., Mishra, V., Hernandez, D., & Gibli, J. (2021). Predicting Neighborhood Change Using Publicly Available Data and Machine Learning. Available at SSRN 3911354.

[24] Chen, James. “Real Estate: Definition, Types, How to Invest in It.” Investopedia, 25 Apr. 2023, www.investopedia.com/terms/r/realestate.asp.

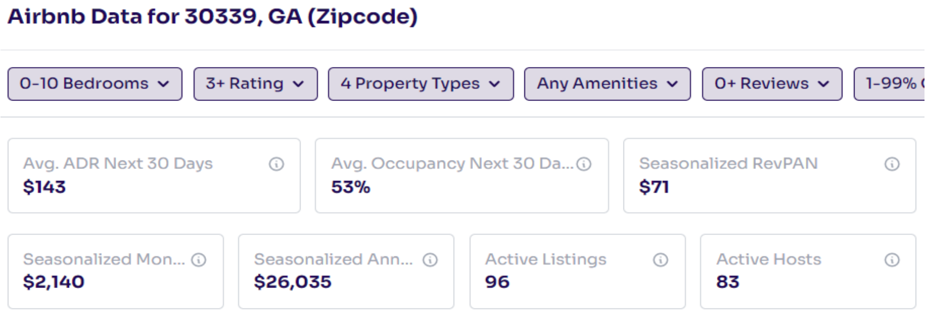
[25] Nguyen, Joseph. “4 Key Factors That Drive the Real Estate Market.” Investopedia, 13 June 2023, www.investopedia.com/articles/mortages-real-estate/11/factors-affecting-real-estate-market.asp.

[26] Mwiti, Derrick. “LightGBM: A Highly-Efficient Gradient Boosting Decision Tree.” KDnuggets, www.kdnuggets.com/2020/06/lightgbm-gradient-boosting-decision-tree.html. Accessed 14 July 2023.

[27] What Is Boosting?, www.ibm.com/topics/boosting. Accessed 15 July 2023.

[28] Team, T. A. (2023, January 6). Gans for Synthetic Data Generation. Towards AI. https://towardsai.net/p/l/gans-for-synthetic-data-generation

APPENDIX



Airbnb Data for 30339, GA (Zipcode)

0-10 Bedrooms | 3+ Rating | 4 Property Types | Any Amenities | 0+ Reviews | 1-99% (

Avg. ADR Next 30 Days
$143

Avg. Occupancy Next 30 Da...
53%

Seasonalized RevPAN
$71

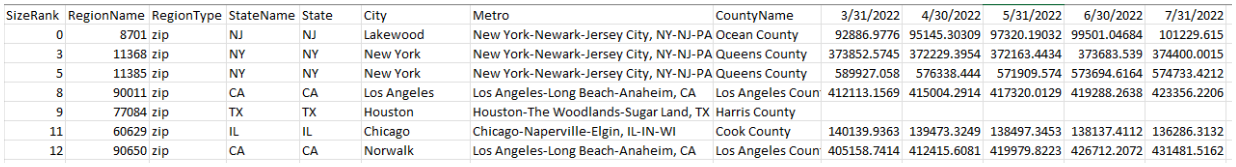
Seasonalized Mon...
$2,140

Seasonalized Ann...
$26,035

Active Listings
96

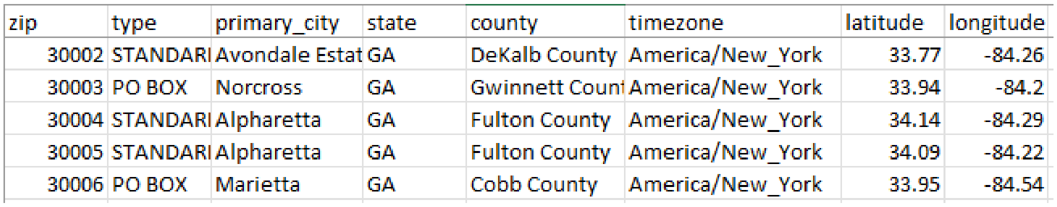
Active Hosts
83

Figure 9: Rabbu STR Market Data. Subset of data fields.



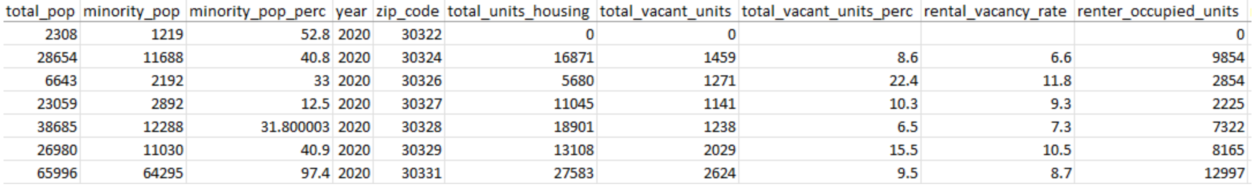
SizeRank	RegionName	RegionType	StateName	State	City	Metro	CountyName	3/31/2022	4/30/2022	5/31/2022	6/30/2022	7/31/2022
0	8701	zip	NJ	NJ	Lakewood	New York-Newark-Jersey City, NY-NJ-PA	Ocean County	92886.9776	95145.30309	97320.19032	99501.04684	101229.615
3	11368	zip	NY	NY	New York	New York-Newark-Jersey City, NY-NJ-PA	Queens County	373852.5745	372229.3954	372163.4434	373683.539	374400.0015
5	11385	zip	NY	NY	New York	New York-Newark-Jersey City, NY-NJ-PA	Queens County	589927.058	576338.444	571909.574	573694.6164	574733.4212
8	90011	zip	CA	CA	Los Angeles	Los Angeles-Long Beach-Anaheim, CA	Los Angeles Coun	412113.1569	415004.2914	417320.0129	419288.2638	423356.2206
9	77084	zip	TX	TX	Houston	Houston-The Woodlands-Sugar Land, TX	Harris County					
11	60629	zip	IL	IL	Chicago	Chicago-Naperville-Elgin, IL-IN-WI	Cook County	140139.9363	139473.3249	138497.3453	138137.4112	136286.3132
12	90650	zip	CA	CA	Norwalk	Los Angeles-Long Beach-Anaheim, CA	Los Angeles Coun	405158.7414	412415.6081	419979.8223	426712.2072	431481.5162

Figure 10: Zillow Home Value Index (ZHVI). Subset of data fields.



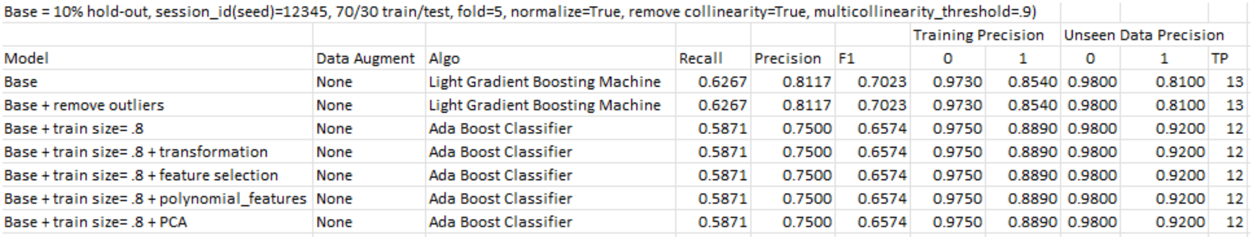
zip	type	primary_city	state	county	timezone	latitude	longitude
30002	STANDARI	Avondale Estat	GA	DeKalb County	America/New_York	33.77	-84.26
30003	PO BOX	Norcross	GA	Gwinnett Count	America/New_York	33.94	-84.2
30004	STANDARI	Alpharetta	GA	Fulton County	America/New_York	34.14	-84.29
30005	STANDARI	Alpharetta	GA	Fulton County	America/New_York	34.09	-84.22
30006	PO BOX	Marietta	GA	Cobb County	America/New_York	33.95	-84.54

Figure 11: Zip Codes Database. Subset of data fields.



total_pop	minority_pop	minority_pop_perc	year	zip_code	total_units_housing	total_vacant_units	total_vacant_units_perc	rental_vacancy_rate	renter_occupied_units
2308	1219	52.8	2020	30322	0	0			0
28654	11688	40.8	2020	30324	16871	1459	8.6	6.6	9854
6643	2192	33	2020	30326	5680	1271	22.4	11.8	2854
23059	2892	12.5	2020	30327	11045	1141	10.3	9.3	2225
38685	12288	31.800003	2020	30328	18901	1238	6.5	7.3	7322
26980	11030	40.9	2020	30329	13108	2029	15.5	10.5	8165
65996	64295	97.4	2020	30331	27583	2624	9.5	8.7	12997

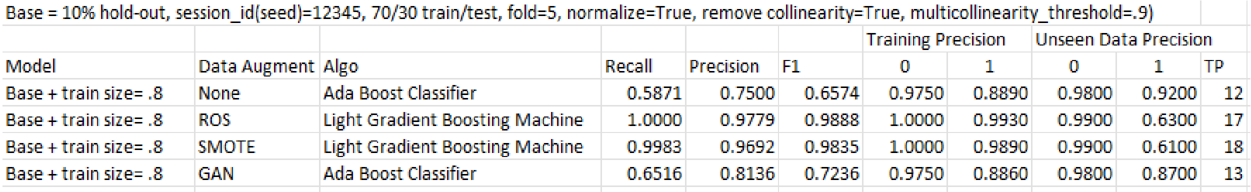
Figure 12: Sample of extracted U.S. Census American Community Survey (ACS) socio-economic data. Data provided by Puy Arena et al. [1].



Base = 10% hold-out, session_id(seed)=12345, 70/30 train/test, fold=5, normalize=True, remove collinearity=True, multicollinearity_threshold=.9)

						Training Precision		Unseen Data Precision		
Model	Data Augment	Algo	Recall	Precision	F1	0	1	0	1	TP
Base	None	Light Gradient Boosting Machine	0.6267	0.8117	0.7023	0.9730	0.8540	0.9800	0.8100	13
Base + remove outliers	None	Light Gradient Boosting Machine	0.6267	0.8117	0.7023	0.9730	0.8540	0.9800	0.8100	13
Base + train size= .8	None	Ada Boost Classifier	0.5871	0.7500	0.6574	0.9750	0.8890	0.9800	0.9200	12
Base + train size= .8 + transformation	None	Ada Boost Classifier	0.5871	0.7500	0.6574	0.9750	0.8890	0.9800	0.9200	12
Base + train size= .8 + feature selection	None	Ada Boost Classifier	0.5871	0.7500	0.6574	0.9750	0.8890	0.9800	0.9200	12
Base + train size= .8 + polynomial_features	None	Ada Boost Classifier	0.5871	0.7500	0.6574	0.9750	0.8890	0.9800	0.9200	12
Base + train size= .8 + PCA	None	Ada Boost Classifier	0.5871	0.7500	0.6574	0.9750	0.8890	0.9800	0.9200	12

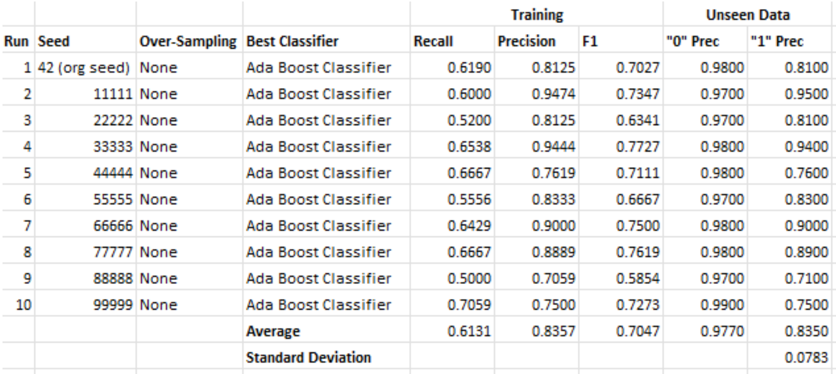
Figure 15: Classification model parameter fine tuning summary results.



Base = 10% hold-out, session_id(seed)=12345, 70/30 train/test, fold=5, normalize=True, remove collinearity=True, multicollinearity_threshold=.9)

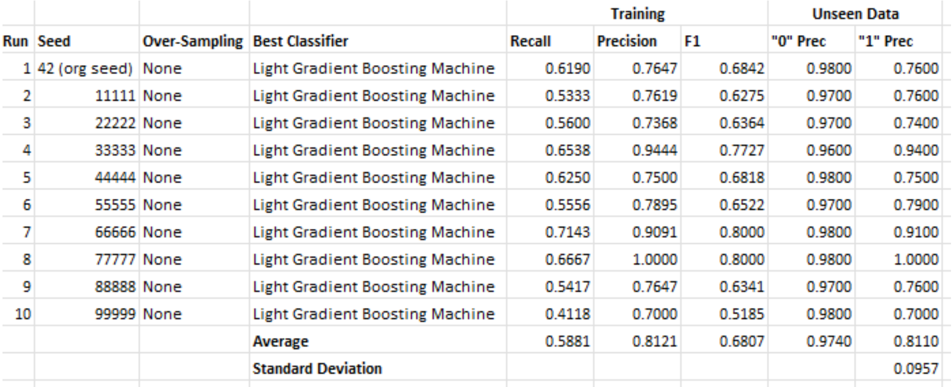
						Training Precision		Unseen Data Precision		
Model	Data Augment	Algo	Recall	Precision	F1	0	1	0	1	TP
Base + train size= .8	None	Ada Boost Classifier	0.5871	0.7500	0.6574	0.9750	0.8890	0.9800	0.9200	12
Base + train size= .8	ROS	Light Gradient Boosting Machine	1.0000	0.9779	0.9888	1.0000	0.9930	0.9900	0.6300	17
Base + train size= .8	SMOTE	Light Gradient Boosting Machine	0.9983	0.9692	0.9835	1.0000	0.9890	0.9900	0.6100	18
Base + train size= .8	GAN	Ada Boost Classifier	0.6516	0.8136	0.7236	0.9750	0.8860	0.9800	0.8700	13

Figure 16: Effect on model performance when using resampled versions of the training data.



				Training			Unseen Data	
Run	**Seed**	**Over-Sampling**	**Best Classifier**	**Recall**	**Precision**	**F1**	**"0" Prec**	**"1" Prec**
1	42 (org seed)	None	Ada Boost Classifier	0.6190	0.8125	0.7027	0.9800	0.8100
2	11111	None	Ada Boost Classifier	0.6000	0.9474	0.7347	0.9700	0.9500
3	22222	None	Ada Boost Classifier	0.5200	0.8125	0.6341	0.9700	0.8100
4	33333	None	Ada Boost Classifier	0.6538	0.9444	0.7727	0.9800	0.9400
5	44444	None	Ada Boost Classifier	0.6667	0.7619	0.7111	0.9800	0.7600
6	55555	None	Ada Boost Classifier	0.5556	0.8333	0.6667	0.9700	0.8300
7	66666	None	Ada Boost Classifier	0.6429	0.9000	0.7500	0.9800	0.9000
8	77777	None	Ada Boost Classifier	0.6667	0.8889	0.7619	0.9800	0.8900
9	88888	None	Ada Boost Classifier	0.5000	0.7059	0.5854	0.9700	0.7100
10	99999	None	Ada Boost Classifier	0.7059	0.7500	0.7273	0.9900	0.7500
			Average	0.6131	0.8357	0.7047	0.9770	0.8350
			Standard Deviation					0.0783

Figure 17: Overfitting test results for the Ada Boost Classifier model.



				Training			Unseen Data	
Run	**Seed**	**Over-Sampling**	**Best Classifier**	**Recall**	**Precision**	**F1**	**"0" Prec**	**"1" Prec**
1	42 (org seed)	None	Light Gradient Boosting Machine	0.6190	0.7647	0.6842	0.9800	0.7600
2	11111	None	Light Gradient Boosting Machine	0.5333	0.7619	0.6275	0.9700	0.7600
3	22222	None	Light Gradient Boosting Machine	0.5600	0.7368	0.6364	0.9700	0.7400
4	33333	None	Light Gradient Boosting Machine	0.6538	0.9444	0.7727	0.9600	0.9400
5	44444	None	Light Gradient Boosting Machine	0.6250	0.7500	0.6818	0.9800	0.7500
6	55555	None	Light Gradient Boosting Machine	0.5556	0.7895	0.6522	0.9700	0.7900
7	66666	None	Light Gradient Boosting Machine	0.7143	0.9091	0.8000	0.9800	0.9100
8	77777	None	Light Gradient Boosting Machine	0.6667	1.0000	0.8000	0.9800	1.0000
9	88888	None	Light Gradient Boosting Machine	0.5417	0.7647	0.6341	0.9700	0.7600
10	99999	None	Light Gradient Boosting Machine	0.4118	0.7000	0.5185	0.9800	0.7000
			Average	0.5881	0.8121	0.6807	0.9740	0.8110
			Standard Deviation					0.0957

Figure 18: Overfitting test results for the LGBM Classifier model.