Research Position Whitepaper: A Survey of Machine Learning Techniques Applied to Real Estate Predictive Analytics

v1.2

Nov 3, 2023

TABLE OF CONTENTS

Overview

In the dynamic terrain of real estate, the ability to rapidly identify properties primed for listing, assess contactability, and gauge seller openness to agent engagement stands as a competitive frontier. The application of sophisticated machine learning (ML) techniques to parse and predict these market signals has opened a new chapter in real estate analytics. This white paper delves into the innovative intersection of ML algorithms and real estate listings, providing a comprehensive examination of how predictive models can be harnessed to sift through vast datasets and discern patterns indicative of a property's likelihood to list, homeowner's receptiveness to phone contact, and propensity to engage in substantive dialogue with real estate professionals. In addition to these predictive models, we explore related applications of intelligence to discern similarity amongst homes and the estimation of market value. Our analysis elucidates the methodologies that can transform raw data into actionable insights, thereby enabling agents to optimize their lead generation processes with unprecedented precision and efficiency.

Background: Types of Models

Statistical classification algorithms are a cornerstone of machine learning, designed to accurately categorize data into different classes. ArchAgent continues to evaluate sophisticated classification techniques in order to find those that work best for predicting specific outcomes. These techniques include:

(i) Gradient Boost Machines (GBM): GBMs are a type of ensemble learning method, where multiple models are trained to solve the same problem and combined to get better results. Specifically, GBMs use boosting, which involves training models sequentially. Each new model focuses on and learns from the errors of the previous ones in an effort to correct them. Gradient boosting involves the use of a gradient descent algorithm to minimize the loss when adding new models. This technique is particularly powerful for predictive tasks where complex patterns are present in the data.

(ii) Random Forest: Random Forest is another ensemble learning method that operates by constructing a multitude of decision trees during training time and outputting the class that is the mode of the classes of the individual trees. It introduces randomness by selecting a subset of the features at each split in the learning process, which helps in making the model more robust to noise and preventing overfitting. Random Forest can handle a large dataset with higher dimensionality and can estimate which variables are important in the classification.

(iii) Radial Basis Function Neural Network (RBF): Radial Basis Function Neural Network is a type of artificial neural network that uses radial basis functions as activation functions. It is primarily used for function approximation, interpolation, and classification problems. The network consists of an input layer, a single hidden layer where the computation with RBFs takes place, and an output layer. The hidden layer of an RBF Network consists of neurons equipped with a radial basis function, most commonly a Gaussian function. Each neuron in the hidden layer acts as a detector that fires strongly when the input is close to a specific point (the center of the RBF), with the response diminishing as the input moves away from the center; this is the radial aspect of the function. The RBF Neural Network can be related to kernel density estimation in the way it uses a similar approach for function approximation. Kernel density estimation is a non-parametric way to estimate the probability density function of a random variable. It places a kernel function (like a Gaussian) on each data point in the dataset and adds up the contributions from each kernel function to estimate the density. In the context of an RBF Neural Network, each neuron can be seen as placing a kernel over the input space, and the output is a weighted sum of these kernels. The parameters of the RBFs (such as the center and width of the Gaussian functions) are adjusted during the training process to fit the data. The network learns to approximate the density of the inputs by positioning and scaling the RBFs appropriately, thereby modeling the data distribution. In essence, while kernel density estimation is used for estimating the probability distribution of data, an RBF Neural Network can be viewed as using a similar principle for more general purposes of approximation and interpolation, effectively capturing the underlying trends or patterns in the data. RBF Neural Networks can be helpful as an alternative to time series when predicting recurring events such as contactability or engagement peaks (i.e., interest) and values (i.e., fatigue/over-contacting.)

(iv) Support Vector Machines (SVMs): Support Vector Machines are a set of supervised learning methods used for classification, regression, and outlier detection. The strength of SVMs lies in their ability to perform well in high-dimensional spaces, even in cases where the number of dimensions exceeds the number of samples. SVMs are particularly adept at binary classification problems. They work by finding the hyperplane that best divides a dataset into two classes, with the goal of maximizing the margin between the closest points of the classes, which are known as support vectors. This creates as wide a gap as possible, allowing for new data to be classified with more confidence. SVMs are versatile in that they can accommodate linear and non-linear boundaries between classes, thanks to the kernel trick, which implicitly maps their inputs into high-dimensional feature spaces. This makes SVMs a powerful tool for a variety of complex classification tasks in fields such as image recognition, bioinformatics, and text categorization, where the boundary between different classes is not immediately clear and the dimensionality of the data is substantial.

(v) Stacking Ensemble/Ensemble Methods: Stacking Ensemble and ensemble methods (or stacked generalization) involve combining multiple different models in a two-or-more level structure. The first level consists of different base classifiers whose predictions are combined and used as input to a second-level model, often called a meta-classifier, which makes the final prediction (or passes along to the next level, etc.) This approach leverages the strengths of each individual classifier and can often achieve better performance than any single classifier alone. It is a more sophisticated technique that can handle a variety of data types and distributions.

ArchAgent has researched and developed a flexible machine-learning workflow and pipelining architecture to easily enable the training and testing of each of these aforementioned algorithms as well as those that materialize in the future. This capability is an enabler for ArchAgent to determine the efficiency and performance of the various algorithmic approaches and to deploy improvements as more capable algorithms emerge. Each of the discussed algorithms has its strengths and is suited to different types of classification problems. The choice of algorithm often depends on the size, quality, and nature of the data, the task at hand, and the computational resources available. ArchAgent has found that the application of several of the above techniques (GBMs, RF, RBFNN, Stacking Ensembles, etc.) provides the best predictive capability rather than a singular, uniform algorithmic approach.

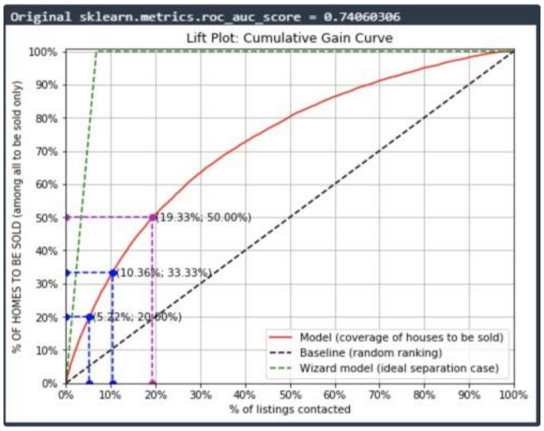
However, the efficacy of a classification model in predicting outcomes is not just intricately tied to the caliber of the underlying algorithms. It is additionally, and crucially, tied to the quality and quantity of the data used in training and testing phases. High-quality real estate data—accurate, complete, and relevant—is the linchpin of any robust predictive model, as it provides the factual groundwork upon which the algorithm learns and hones its predictive capabilities. Similarly, the volume of data plays a significant role: ample data points offer a more comprehensive view of the problem space, allowing the model to discern and generalize patterns effectively. However, simply having a large amount of data is not sufficient if it doesn't capture the variability and complexity of real-world scenarios the model aims to address. It's this intricate dance between sophisticated algorithms and rich datasets that propels a classification model from mere adequacy to excellence, enabling it to yield reliable and insightful predictions that can stand the test of deployment in varied and dynamic environments. ArchAgent has accumulated a wealth of differentiated property, listing, financial, behavioral, household, personal profile, and demographic data over the past 12 years, and selectively with this knowledge, learned to prepare and engineer data for each specific modeling use.

Lift Curves as a Mechanism of Performance Efficiency

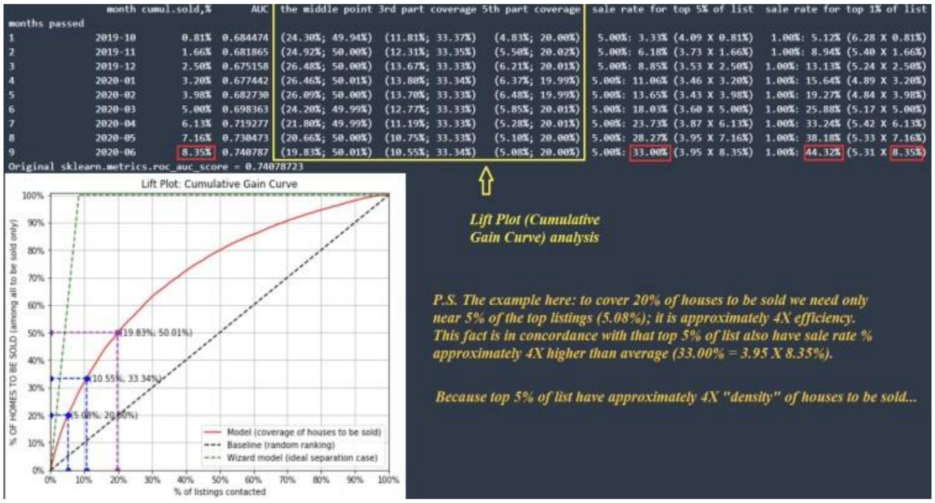
Statistical regression analysis and classification models sit at the heart of predictive analytics, providing powerful tools to forecast future events based on historical data. In the realm of real estate, these methods have been intricately applied to calculate the "propensity to list," which is the likelihood of a homeowner deciding to sell their property and consequently put it on market. The objective of such models is not merely to predict but to do so with remarkable efficiency; for example, aiming to identify a subset of properties—say, 30% of the total—that are highly likely to encompass 90% of all listings that will emerge in the market within the next nine months. This concentration of prediction illustrates an adept use of a lift curve, a tool used to measure the performance of classification models at different threshold settings. In the case of real estate listings, a lift curve with a value of 3x means that the model is three times as effective at identifying potential listings as random selection would be, thereby enabling real estate professionals to target their marketing efforts more precisely and with greater return on investment.

ArchAgent has been able to achieve lift equivalent to that demonstrated by Figure 1 below, a curve plot for the “Southern Stable Metropolis” city type (see below). ArchAgent is able to pack 20% of the sales into only 5.08% and 5.22% of the listings, with efficiency gains of 383% and 395%. Providing this level of predictability to real estate agents increases their productivity by 3.83-3.95x.

Figure 1: Lift Curves for Southern Stable and Stable Large Metropolises



```
Original sklearn.metrics.roc_auc_score = 0.74060306
```



```
                month cumul.sold,%       AUC  the middle point 3rd part coverage 5th part coverage  sale rate for top 5% of list  sale rate for top 1% of list
months passed
1               2019-10         0.81%  0.684474  (24.30%; 49.94%)  (11.81%; 33.37%)   (4.83%; 20.00%)   5.00%: 3.33% (4.09 X 0.81%)   1.00%: 5.12% (6.28 X 0.81%)
2               2019-11         1.66%  0.681865  (24.92%; 50.00%)  (12.31%; 33.35%)   (5.50%; 20.02%)   5.00%: 6.18% (3.73 X 1.66%)   1.00%: 8.94% (5.40 X 1.66%)
3               2019-12         2.50%  0.675158  (26.48%; 50.00%)  (13.67%; 33.33%)   (6.21%; 20.01%)   5.00%: 8.85% (3.53 X 2.50%)  1.00%: 13.13% (5.24 X 2.50%)
4               2020-01         3.20%  0.677442  (26.46%; 50.01%)  (13.80%; 33.34%)   (6.37%; 19.99%)  5.00%: 11.06% (3.46 X 3.20%)  1.00%: 15.64% (4.89 X 3.20%)
5               2020-02         3.98%  0.682730  (26.09%; 50.00%)  (13.70%; 33.33%)   (6.48%; 19.99%)  5.00%: 13.65% (3.43 X 3.98%)  1.00%: 19.27% (4.84 X 3.98%)
6               2020-03         5.00%  0.698363  (24.20%; 49.99%)  (12.77%; 33.33%)   (5.85%; 20.01%)  5.00%: 18.03% (3.60 X 5.00%)  1.00%: 25.88% (5.17 X 5.00%)
7               2020-04         6.13%  0.719277  (21.80%; 49.99%)  (11.19%; 33.33%)   (5.28%; 20.01%)  5.00%: 23.73% (3.87 X 6.13%)  1.00%: 33.24% (5.42 X 6.13%)
8               2020-05         7.16%  0.730473  (20.66%; 50.00%)  (10.75%; 33.33%)   (5.10%; 20.00%)  5.00%: 28.27% (3.95 X 7.16%)  1.00%: 38.18% (5.33 X 7.16%)
9               2020-06         8.35%  0.740787  (19.83%; 50.01%)  (10.55%; 33.34%)   (5.08%; 20.00%)  5.00%: 33.00% (3.95 X 8.35%)  1.00%: 44.32% (5.31 X 8.35%)
Original sklearn.metrics.roc_auc_score = 0.74078723
```

Lift Plot: Cumulative Gain Curve
% OF HOMES TO BE SOLD (among all to be sold only)
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
(19.83%; 50.01%)
(10.55%; 33.34%)
(5.08%; 20.00%)
Model (coverage of houses to be sold)
Baseline (random ranking)
Wizard model (ideal separation case)
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
% of listings contacted

Lift Plot (Cumulative Gain Curve) analysis

P.S. The example here: to cover 20% of houses to be sold we need only near 5% of the top listings (5.08%); it is approximately 4X efficiency. This fact is in concordance with that top 5% of list also have sale rate % approximately 4X higher than average (33.00% = 3.95 X 8.35%).

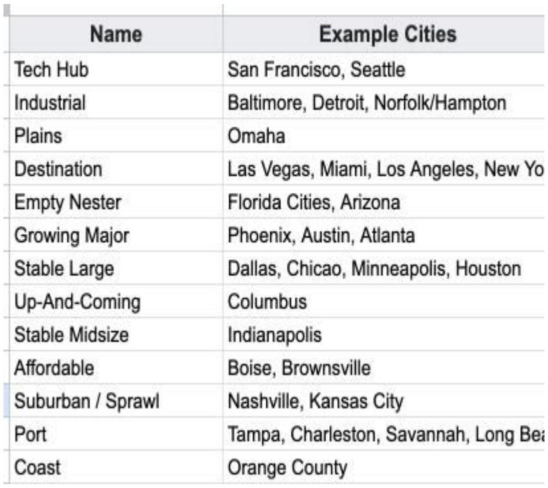
Because top 5% of list have approximately 4X "density" of houses to be sold...

Propensity-to-List Model Types and Spatial (Geographic) Prediction Models

Another technique that ArchAgent has used to improve propensity-to-list model performance is to create a multitude of spatial prediction models that cover representative microeconomic real estate markets. These models represent different canonical, repetitive trends in real estate that cover market spectrums like sales rate (cold to hot), housing age (old to new), population (stagnant to growing), location (rural to urban), available inventory (low to high), etc. These models take into account the old adage that the three most important factors in real estate are "location, location, location." They leverage spatial and geographic property attributes and relationships, recognizing that the value and sales potential of a real estate asset   are significantly influenced by its location and the characteristics of its surrounding area. Key factors include local sales and inventory metrics; proximity to amenities such as schools, parks, and shopping centers; accessibility to major roadways and public transportation; the gentrification of a neighborhood or the expansion of urban sprawl; and other assorted neighborhood demographics (even the incidence of natural light or the prevalence of certain architectural styles can influence propensity-to-list models.) In the age of big data, the incorporation of as many data streams and signals as possible allows ArchAgent to gain a dynamic and predictive understanding of real estate markets at a granular level.

Figure 2 below is an example of canonical spatial models that ArchAgent has created. ArchAgent tests granular geographic regions (no larger than single zip code boundaries) against each of its spatial prediction models in order to understand what particular model has the best predictive performance for a specific geographic hyper-locale.

Figure 2: Sample Canonical Spatial Prediction Models

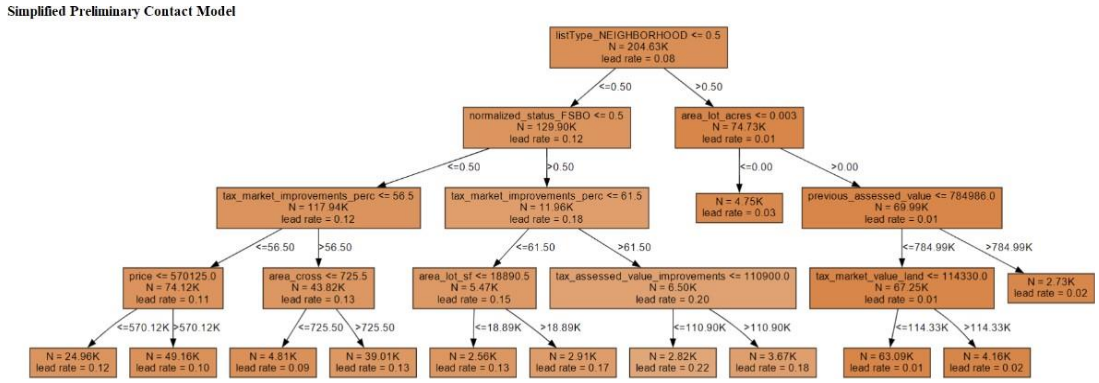


Name	Example Cities
Tech Hub	San Francisco, Seattle
Industrial	Baltimore, Detroit, Norfolk/Hampton
Plains	Omaha
Destination	Las Vegas, Miami, Los Angeles, New Yo
Empty Nester	Florida Cities, Arizona
Growing Major	Phoenix, Austin, Atlanta
Stable Large	Dallas, Chicao, Minneapolis, Houston
Up-And-Coming	Columbus
Stable Midsize	Indianapolis
Affordable	Boise, Brownsville
Suburban / Sprawl	Nashville, Kansas City
Port	Tampa, Charleston, Savannah, Long Bea
Coast	Orange County

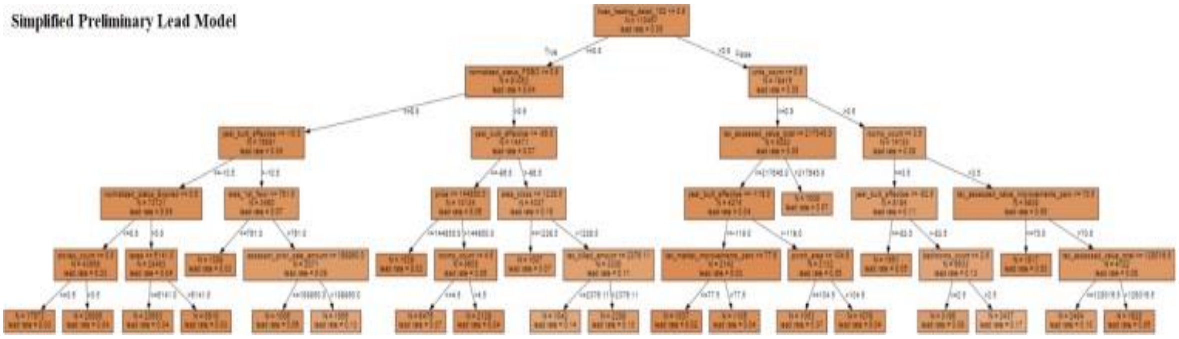
Contactability and Engagement Modeling

In addition to predicting propensity-to-list, ArchAgent has studied and modeled which outreach behaviors by real estate agents achieve intended outcomes. These models predict likely to contact and likely to engage (i.e., result in a real estate agent follow-up or lead). Two unique phenomena are of special consideration and deserve special attention related to contact and engagement modeling. First, the process of prediction includes two different steps: a) classification modeling to achieve efficient, high-performance predictions; and b) visualization modeling to provide interpretable and understandable outcomes. While the same statistical regression techniques and classification models that apply to propensity-to-list also apply to contactability and engagement classification models, ArchAgent uses decision trees (Figure 3) to provide more digestible and understandable predictions for our users.

Figure 3: Contact and Lead Model Decision Tree Visualizations

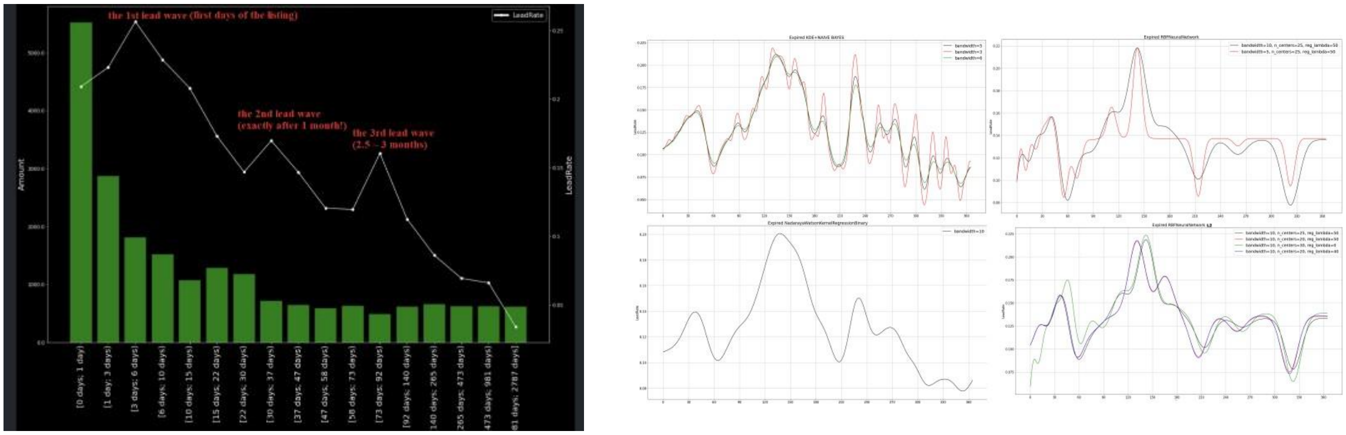


The information contained in this document is considered privileged, confidential, and protected from disclosure. If the viewer of this document is not the original recipient, you are hereby notified that any dissemination, distribution, copying, or other use of this communication is strictly prohibited. All rights reserved. Confidential and Proprietary.



Second, ArchAgent applies regular user feedback and real-time disposition information to further improve predictions. Business rules, based on domain expertise, enhance the accuracy of these predictions by considering aggregate contact rates and user dispositions. Decay strategies factor in unsuccessful attempts for likely-to-contact and negative disposition outcomes for likely-to-lead to gradually reduce previous prediction scores, account for changing trends, and diminish the impact of outdated information as time progresses. Implementing decay strategies helps the models adapt to the most current patterns and reduces the risk of overfitting to historical data that no longer reflects the present reality, thus maintaining the model's accuracy and robustness over time. In a similar manner, as contactability and engagement is cyclical (i.e., those interested in selling their home may have peak-and-valley type reactions to being approach over certain periods of time), ArchAgent models responsiveness cycles to understand what time intervals demonstrate for certain property owners their interest in engaging with real estate agents who are contacting during first-approach cycles, second approaches, and subsequent later approaches in time (see Figure 4).

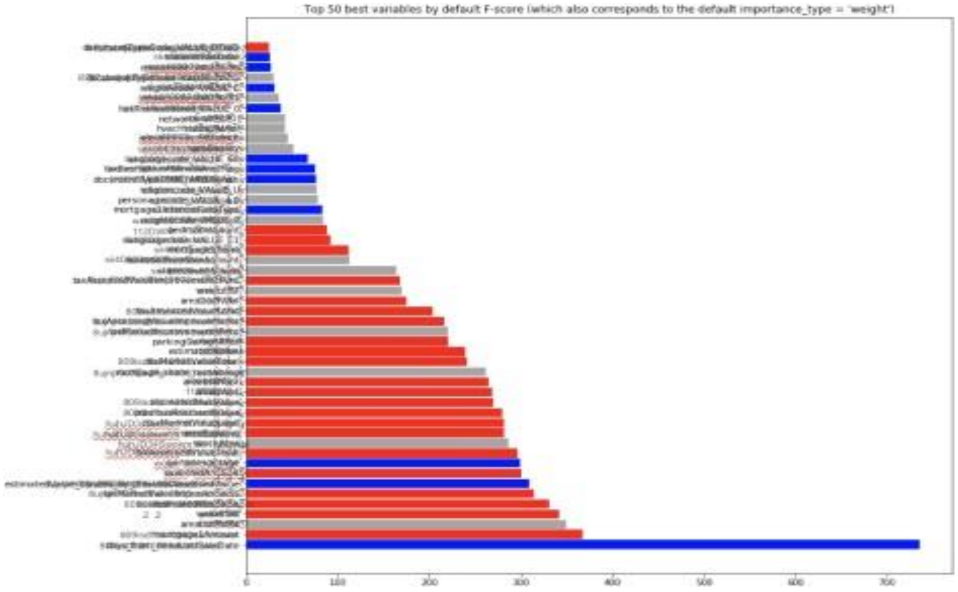
Figure 4: Lead Rates Over Time / Kernel Density & Support Vector Regression (and other technique) Estimators



Propensity-to-List Variable Significance and Feature Engineering

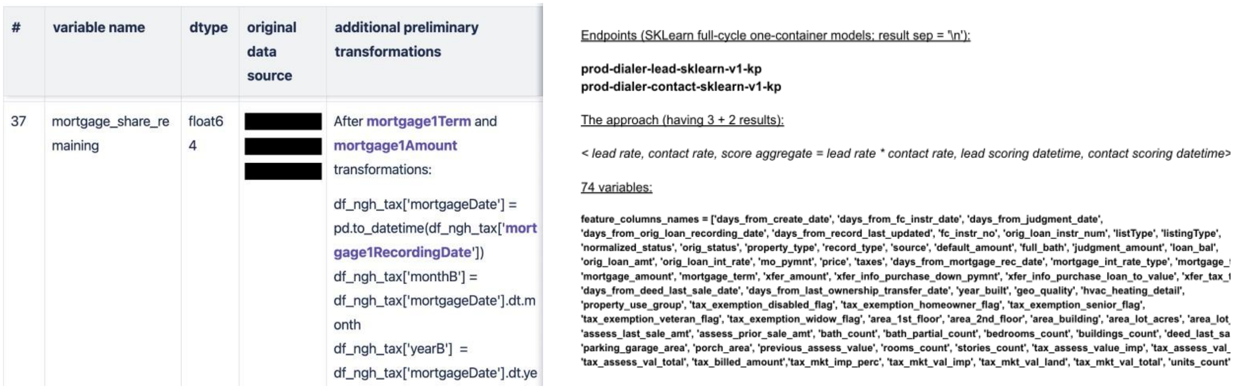
In addition to scoring properties for propensity-to-list, contactability, and engagement, ArchAgent has   invested considerable research into explaining to users the outcomes of models and which characteristics correlate with desired outcomes. As a part of this process, ArchAgent performs variable significance analysis for each-and-every predictive model it generates. Typically, the top 50 variables that both positively and negatively correlate to predicted outcomes are calculated and visualized (Figure 5).

Figure 5: Obfuscated Variable Names Significant Variable Importance Chart



ArchAgent also researches and identifies which variables may need to be encoded, prepared, and otherwise calculated in order to improve model performance. It is not unusual for ArchAgent to consider hundreds if not thousands of variables in prediction models. Typically, ArchAgent will test both deep and shallow models (i.e., fewer engineered and total features) in order to arrive at a model that provides a balance between compute performance and deployability without compromising model efficacy. ArchAgent has developed ample expertise in this area of R&D achieving model size reductions that have resulted in models with as few as 75-300 input variables and those that contain features engineered specifically for model improvements (Figure 6).

Figure 6: Reduced Model Metadata, Engineered Features Matrix



#	variable name	dtype	original data source	additional preliminary transformations
37	mortgage_share_remaining	float64	[redacted]	After **mortgage1Term** and **mortgage1Amount** transformations: df_ngh_tax['mortgageDate'] = pd.to_datetime(df_ngh_tax['**mortgage1RecordingDate**']) df_ngh_tax['monthB'] = df_ngh_tax['mortgageDate'].dt.month df_ngh_tax['yearB'] = df_ngh_tax['mortgageDate'].dt.ye

Endpoints (SKLearn full-cycle one-container models; result sep = '\n'):

prod-dialer-lead-sklearn-v1-kp
prod-dialer-contact-sklearn-v1-kp

The approach (having 3 + 2 results):

*< lead rate, contact rate, score aggregate = lead rate * contact rate, lead scoring datetime, contact scoring datetime>*

74 variables:

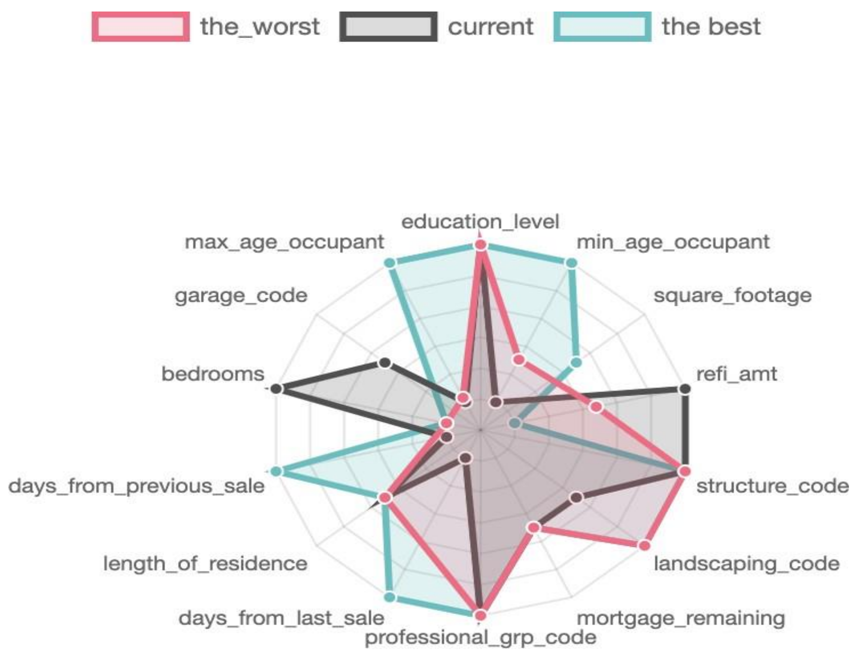
```
feature_columns_names = ['days_from_create_date', 'days_from_fc_instr_date', 'days_from_judgment_date',
'days_from_orig_loan_recording_date', 'days_from_record_last_updated', 'fc_instr_no', 'orig_loan_instr_num', 'listType', 'listingType',
'normalized_status', 'orig_status', 'property_type', 'record_type', 'source', 'default_amount', 'full_bath', 'judgment_amount', 'loan_bal',
'orig_loan_amt', 'orig_loan_int_rate', 'mo_pymnt', 'price', 'taxes', 'days_from_mortgage_rec_date', 'mortgage_int_rate_type', 'mortgage_
'mortgage_amount', 'mortgage_term', 'xfer_amount', 'xfer_info_purchase_down_pymnt', 'xfer_info_purchase_loan_to_value', 'xfer_tax_
'days_from_deed_last_sale_date', 'days_from_last_ownership_transfer_date', 'year_built', 'geo_quality', 'hvac_heating_detail',
'property_use_group', 'tax_exemption_disabled_flag', 'tax_exemption_homeowner_flag', 'tax_exemption_senior_flag',
'tax_exemption_veteran_flag', 'tax_exemption_widow_flag', 'area_1st_floor', 'area_2nd_floor', 'area_building', 'area_lot_acres', 'area_lot_
'assess_last_sale_amt', 'assess_prior_sale_amt', 'bath_count', 'bath_partial_count', 'bedrooms_count', 'buildings_count', 'deed_last_sa
'parking_garage_area', 'porch_area', 'previous_assess_value', 'rooms_count', 'stories_count', 'tax_assess_value_imp', 'tax_assess_val_
'tax_assess_val_total', 'tax_billed_amount','tax_mkt_imp_perc', 'tax_mkt_val_imp', 'tax_mkt_val_land', 'tax_mkt_val_total', 'units_count'
```

Similar Properties and Variable Scores

Real estate properties can be compared to one another by evaluating their similarities and differences across a range of attributes known to influence their propensity-to-list for sale. This process involves identifying key characteristics such as location, property size, age, condition, market trends, and economic indicators that are statistically significant in predicting a property's propensity-to-list. By analyzing how closely the attributes of one property correlate with another, we can judge their similarity in terms of listing behavior. To visualize these comparisons, ArchAgent employs an exceptionally useful visual: spider diagrams, also known as radar charts (Figure 7). A spider diagram plots the values of each attribute on a separate axis that starts from the center of the chart and radiates outward. Each property's attributes are plotted on these axes and then connected to form a polygon. When multiple properties are plotted on the same spider diagram, it becomes easy to see how one property's attributes compare to another by observing the shape and size of their respective polygons. Overlapping areas may indicate a strong correlation in certain attributes, suggesting similar listing propensities, while divergent parts of the polygons highlight differences. Mapping both best and worst properties provides the user with comparative baselines to understand where a property falls in terms of its correlations.

For example, if two properties have similar polygons in a spider diagram, showing that they share commonalities in crucial attributes like square footage, number of bedrooms, and proximity to schools, they can be judged to have a similar propensity to list. Conversely, significant differences in the polygons would indicate a lower correlation in listing propensity. This visual tool allows real estate professionals to quickly assess and compare properties on multiple dimensions, aiding in prospecting decisions and market analysis.

Figure 7: Example Spider Diagram Illustrating Property Mapping to Best/Worst Propensity-To-Lis



The information contained in this document is considered privileged, confidential, and protected from disclosure. If the viewer of this document is not the original recipient, you are hereby notified that any dissemination, distribution, copying, or other use of this communication is strictly prohibited. All rights reserved. Confidential and Proprietary.

Automated Valuation Model

ArchAgent has created an Automated Valuation Model (AVM) to appraise property values rapidly and objectively. By utilizing historical data, property characteristics, and various algorithms, ArchAgent AVM generates estimates of market value for residential properties without the need for a physical inspection or human appraisal. ArchAgent incorporates hedonic pricing models, which isolate the impact of individual factors on property prices, as well as spatial autocorrelation, which accounts for the principle that properties close to one another are more likely to have similar values than those farther apart. Additionally, ArchAgent has applied its spatial prediction regression models to predict prices of unsold properties and to understand the potential impact of new developments on existing property values.

ArchAgent's approach is more comprehensive than standard approaches to AVM because it includes not only a predicted or most likely value but also a range of values to capture the potential variability in a property's value. This is achieved by calculating a maximum and a minimum value alongside the predicted value. The maximum value represents a scenario where all the value-adding attributes of the property and market conditions are considered at their most favorable, while the minimum value accounts for the opposite—a conservative estimate under less favorable conditions.

To determine these ranges, ArchAgent model adds and subtracts a specific increment from the predicted value, which is derived from a statistically calculated range of pricing. This increment is based on factors such as the confidence interval of the valuation model, the volatility of the local real estate market, and the uniqueness of the property's features. The result is a valuation band that gives end-users a sense of the potential high and low ends of a property's market value, in addition to the most likely selling price.
For instance, if the AVM predicts a property's value at $300,000 with a calculated increment of $20,000 based on market data and statistical analysis, the maximum value might be set at $320,000, and the minimum value at $280,000. This range allows users to understand the uncertainty in the valuation and to gauge risk appropriately.

Conclusion

In conclusion, this whitepaper has provided a comprehensive examination of the cutting-edge classification models that leverage both statistical regression techniques and machine learning to revolutionize the real estate industry. We have delved into the intricacies of propensity-to-list models, unpacked the nuances of predicting contactability and engagement, and illuminated the sophistication of automated valuation models. Through meticulous feature significance calculations and the artful craft of feature engineering, we have shown how to fine-tune algorithms for optimal performance. Moreover, our exploration of various testing and training methodologies underscores the importance of rigorous validation to ensure reliable predictions. The visualization of results, particularly through spider diagrams, has offered an intuitive understanding of complex data, fostering greater transparency and interpretability. Finally, the strategic application of decay strategies demonstrates a forward-thinking approach to maintaining the relevancy and accuracy of models over time. As the real estate industry continues to evolve, the insights gleaned from this whitepaper will undoubtedly serve as a cornerstone for future innovations, enabling professionals to make data-driven decisions with unprecedented precision and insight.