# Machine Learning for Realtors

Identifying high propensity customers to buy/lease/rent & renew properties

WHITEPAPER

# Table of contents



# Executive Summary

Changing consumer behavior is a perennial problem for businesses of all types and sizes. Real estate businesses are no exception to this. Gone are the days where traditional demographic personas defined the consumers, millennials are becoming less and less predictable with just the superficial demographic data points. Businesses have to dig deeper usually to understand them better, to be able to segment them or to make any meaningful predictions about their buying patterns or behavior. This has prompted Realtors to increasingly employ data from behavioral and financial profiles, along with demographic data to enable a much deeper understanding of the customer journey.

This whitepaper attempts to explain briefly, how machine learning can be used to derive customer insights and how these insights can help Realtors in targeting high potential customers with personalized communication.

## Lead Authors

Manas Pant Manager
L Antony Shajin Sr. Software Engineer

## Contributors

P Prathyusha Data Scientist
Sahitesh Reddypalli Data Scientist
Sreeja Yalamaddi Data Scientist
L Akshay Chandra Data Scientist

## About ACS Solutions

ACS Solutions is a leading global information technology services and consulting organization that has been serving businesses globally across industries since 1998. A trusted partner to both mid-market and Fortune 500 clients, ACS Solutions has been instrumental in each of their unique digital transformation journeys. Our extensive industry-specific domain expertise and passion for innovation has helped clients envision, build, scale and run their businesses more efficiently, for over two decades.

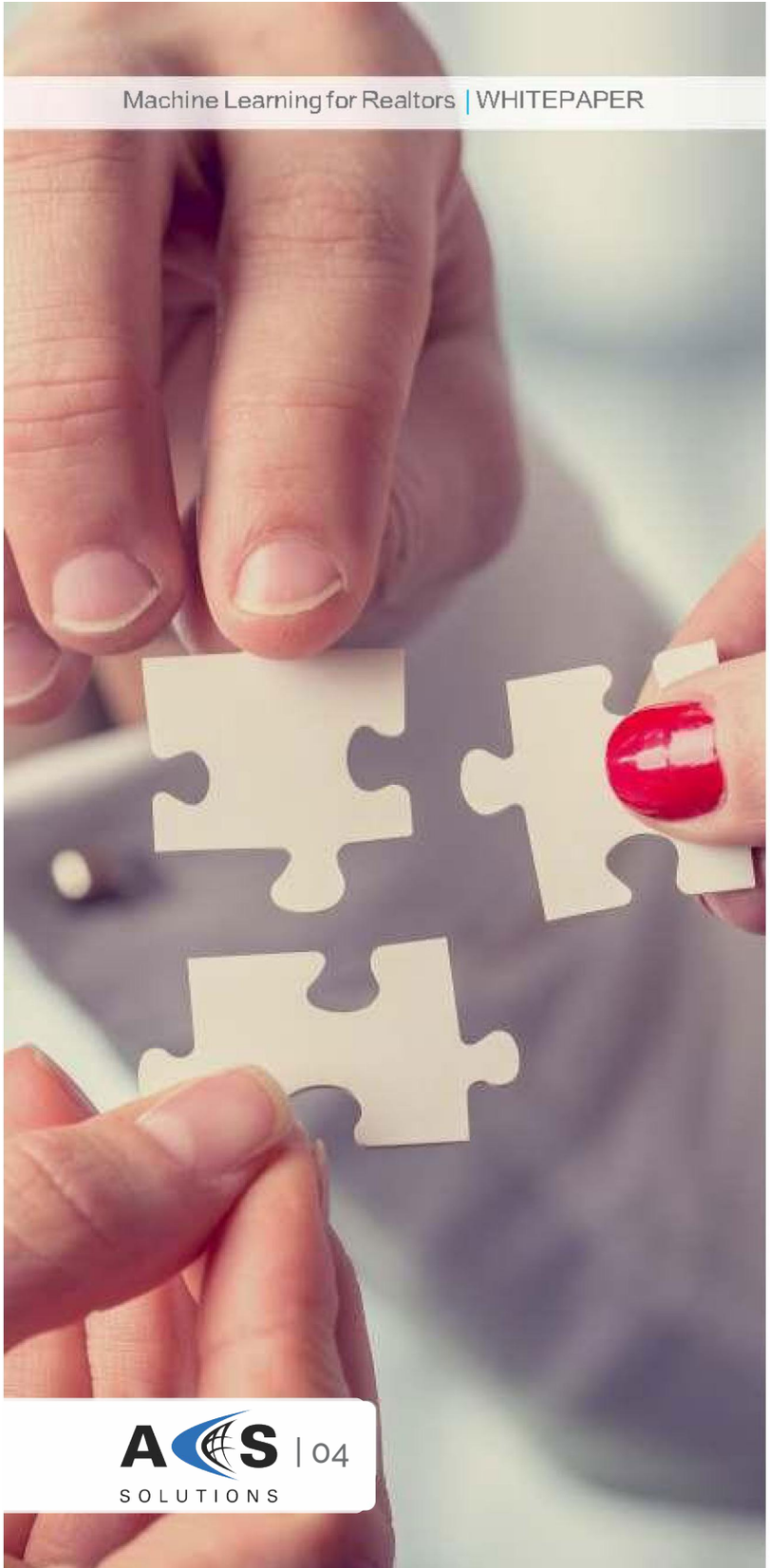
We have a proven track record of developing large and complex software and technology solutions for Fortune 500 clients such as Microsoft, Blue Cross Blue Shield, Cox Communications, and Novartis. We deliver on our commitments and enable our customers to achieve a digital competitive advantage through flexible and global delivery models, agile methodologies, and expert frameworks. Headquartered in Duluth, GA, and with several locations across North America, Europe, and the Asia-Pacific regions, ACS Solutions specializes in 360-degree digital transformation and IT consulting services.

For more information, please reach us at –
acssolutions@acsicorp.com

## Introduction

In real estate advisory services, identifying interested and relevant customers to buy/sell/lease/rent/break-lease/renew properties becomes increasingly important. Identifying high potential customers helps real estate agents to target those specific customers or groups of customers with personalized communications and offers. While real estate agents collect basic information (e.g. demographic, bought or not bought) from their customers/leads, this data alone is not sufficient to analyze their customers' behavior, calculate propensity to buy/sell and perform targeted marketing. Third parties like Experian provide consumer data to learn more about customers' behavioral, psychographic, financial profiles. Blending real estate agent owned information with third party data could bring new business and drive more intelligent interactions with customers.

owned information with third party data could bring new business and drive more intelligent interactions with customers





## Our Approach

We designed and developed solutions to help realtors or real estate agents in the following areas

- Identify higher propensity customers and leads
- Analyze customer groups
- Build communication strategies to approach customers

Our solution includes three machine learning accelerators developed using large volume of consumer marketing data to predict propensity for customers to buy or lease properties. We used algorithms such as Random Forest, Logistic Regression and Support Vector Machines to build the accelerators. We selected the best machine-learning model, which presented highest prediction performance during evaluation and used it for final prediction.

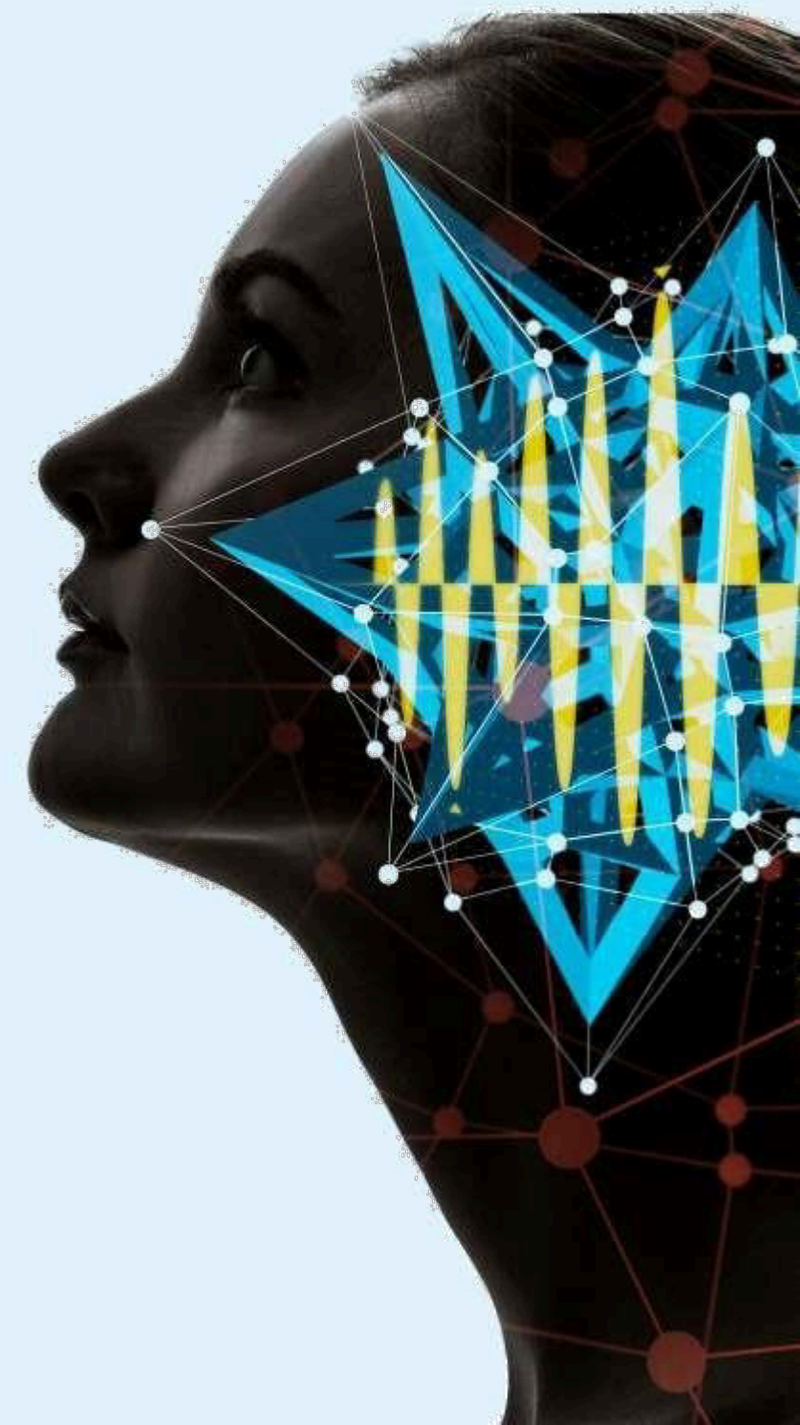
We then designed three solutions to analyze different customer's characteristics. Our solution include self-service machine learning analytics in order to give more control to realtor therefore their domain experts can leverage self-service analytics' power and design recommendations on how to engage effectively with each persona type, thus enabling a superior customer experience.

- Pre-built machine learning analytics
- Self-service machine learning analytics
- Enhanced self-services machine learning analytics

**Pre-built machine learning analytics:** : Pre-built machine learning analytics helps realtors to quickly visualize and analyze different groups'/segments' characteristics on their whole customer population. We selected features using domain knowledge and various machine learning techniques, and built machine-learning accelerators on the pre- selected features for each line of business. E.g. Buy, Lease, Break-Lease and Renew. The realtor needs to choose their line of business. The analytics service then presents realtor with quick cluster analysis report and propensity value for each customer to buy or lease.

**Self-service machine learning analytics:** Self-service machine learning helps realtors visualize and analyze different groups/segments' characteristics on the specific set of customer population they are interested in. This requires realtor to have deeper understanding of domain to choose more relevant features for their line of business.

The self-service machine learning gives control to realtors over the large list features and type of customers that they want to use for machine learning analytics. Realtors can choose features from the list of features using their domain knowledge. It also allows realtors to apply various filters to analyze subsets of customers. The self-service machine learning analytics system then applies machine-learning techniques on the subset of data and the features selected by the realtor. Finally, it presents realtors with customer analysis reports and propensity value for each customer to buy or lease.




**es machine learning analytics:** Enhanced self- service machine
lps realtors visualize and analyze different groups/segments'
e specific type of customer population they are interested in.
require realtor to have in depth domain knowledge to be able
or machine analytics as we already pre-identified features for
. Important features are pre-identified for specific line of
lease using domain knowledge and machine learning

ervice machine learning provides control to realtor over features
ch line of business and type of customers that they want to use
ine-learning analytics. Realtors can choose features from the
res. It also allows them to select subset of customers by applying
nhanced self-service analytics system then applies machine
bset of data and the features selected, and finally presents them
sis reports and propensity value for each customer to buy or

egrated with a web application. The application UI provides the
ML accelerator model and the data filters. The analysis reports
time and are available for download.

## Implementation and Insights

We implemented machine-learning pipeline, which has the stages listed below. The workability of each stage varies with the type of accelerator chosen,

**Data Analysis:** Our solution needs two datasets, third party data that has all the financial, demographic, geo-demographic, psychographic and behavioral characteristics of all customers and realtor's acquired customer data. The third party data is considered as primary dataset for our analysis. We performed Exploratory Data Analysis on these datasets to understand the features and added new features wherever applicable. In case of Enhanced self-service analytics, the filters selected by realtor are also applied on the datasets.

**Below are few insights from Analysis:**

- Some features Like Year Built, Visit Month, etc. have an impact on client to lease the property
- Number of visits and leases are more during the summer season

**Data Preprocessing:** After initial data analysis, we retained the relevant columns from both datasets and merged those using common key columns. We then converted the data into usable format for machine learning, removed the duplicates and performed missing value imputation. We reduced number of categories by binning categorical features with too many categories and performed one hot encoding on all categorical features. We filtered out some of numeric features using Pearson correlation.



**Implementation Stages**

Data Analysis

Data Preprocessing

Feature Selection

- Filter-Based Feature Selection
- Wrapper Methods
- Embedded Method

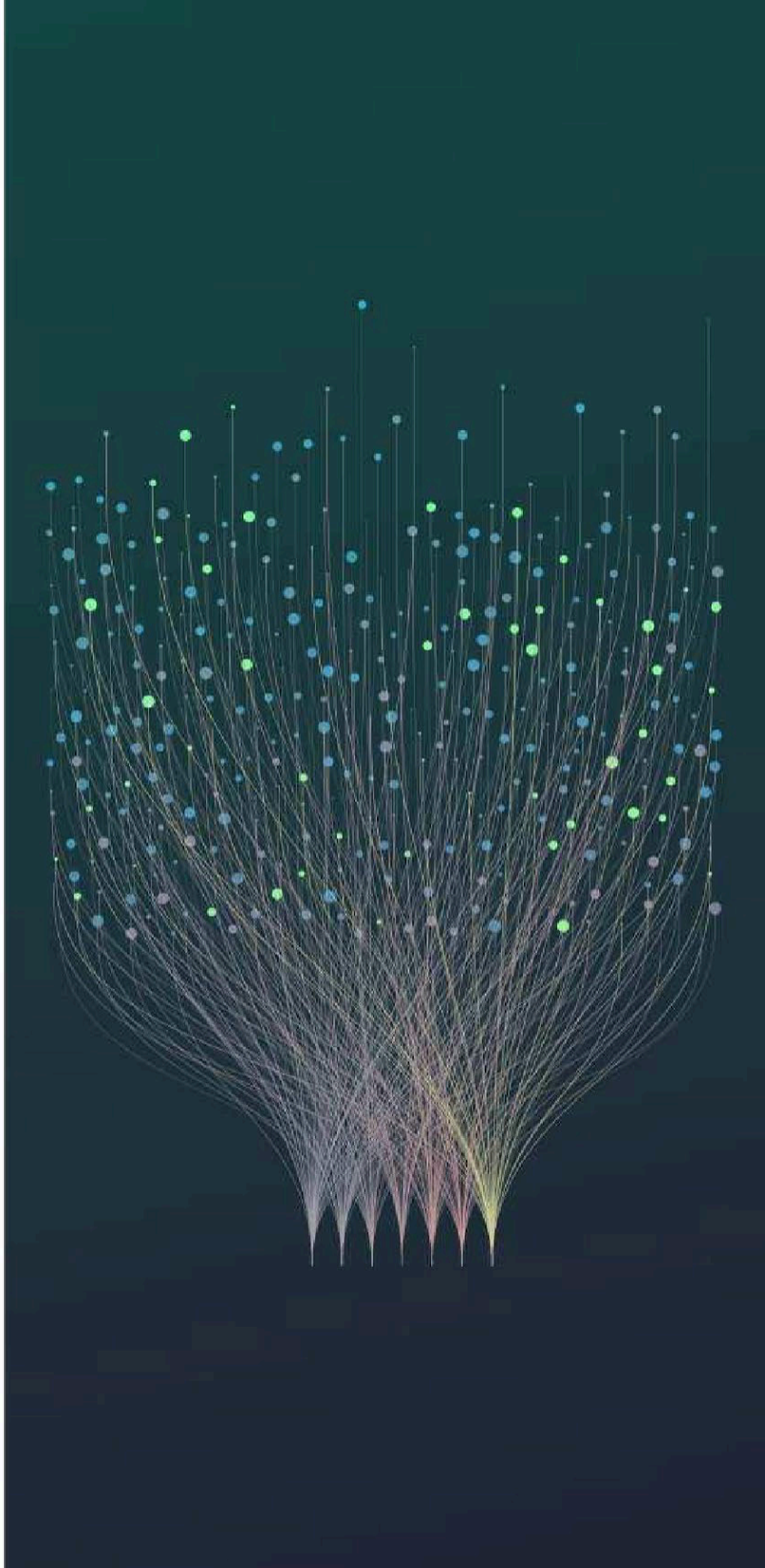
Data Modelling

**Feature Selection:** After pre-processing, we went ahead to select features which are relevant to predict propensity to buy, sell, lease, etc. The dataset had both categorical and continuous features. We followed different ways to get final features list. This step varies from one accelerator to the other.

- In case of self-service machine learning accelerator, the feature set selected by the realtor is directly used for modelling.
- For pre-built machine learning accelerator, the feature selection process is as described below -

  - **Filter-Based Feature Selection:** It provides a selection of widely used statistical tests for determining the subset of input columns that have the greatest predictive power. Here, we perform statistical tests like ANOVA, Chi-Squared, LDA, etc., to get the correlation of features with the output labels. Filter methods work best if features in the data are either all continuous or all categorical. For a mixed set of features, we convert the continuous into categorical by binning which reduces its effectiveness.

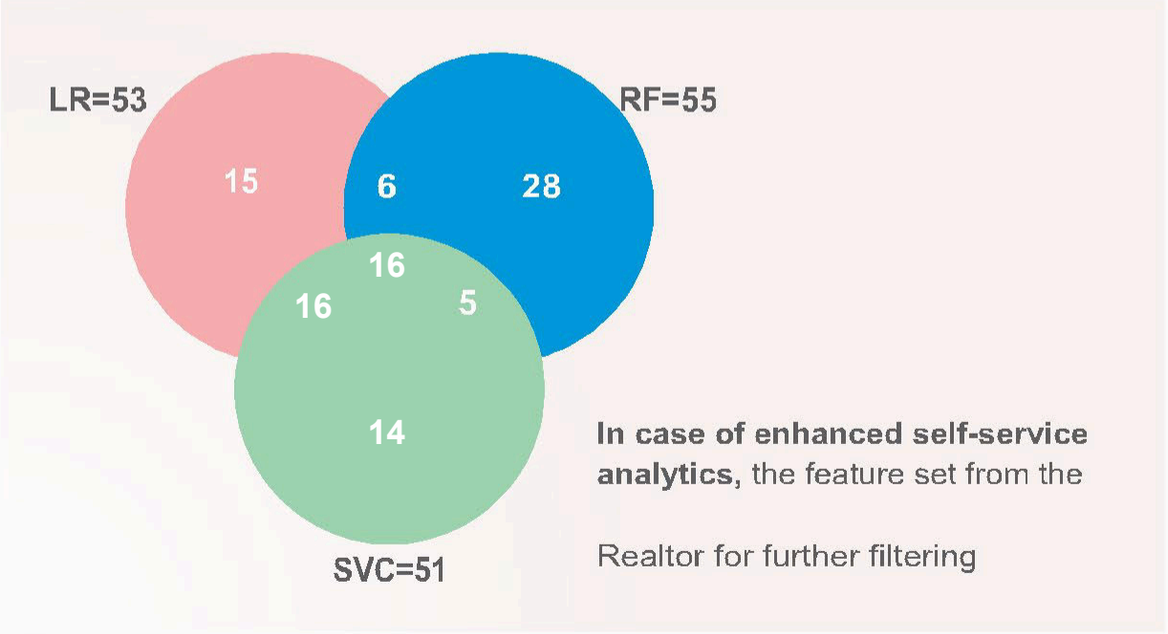
  - **Wrapper Methods:** As the name suggests, it is based on a feedback loop. We recursively run the model with all the features and keep removing least important one till we end up with the best features or start with random set of features and keep adding new features till we get the best model. In addition, they use a classification performance of a classifier (like accuracy, ROC scores) to do the evaluation. Despite being computationally expensive, wrapper based methods are advantageous for giving better performances since they use the target in classifier for the feature selection.

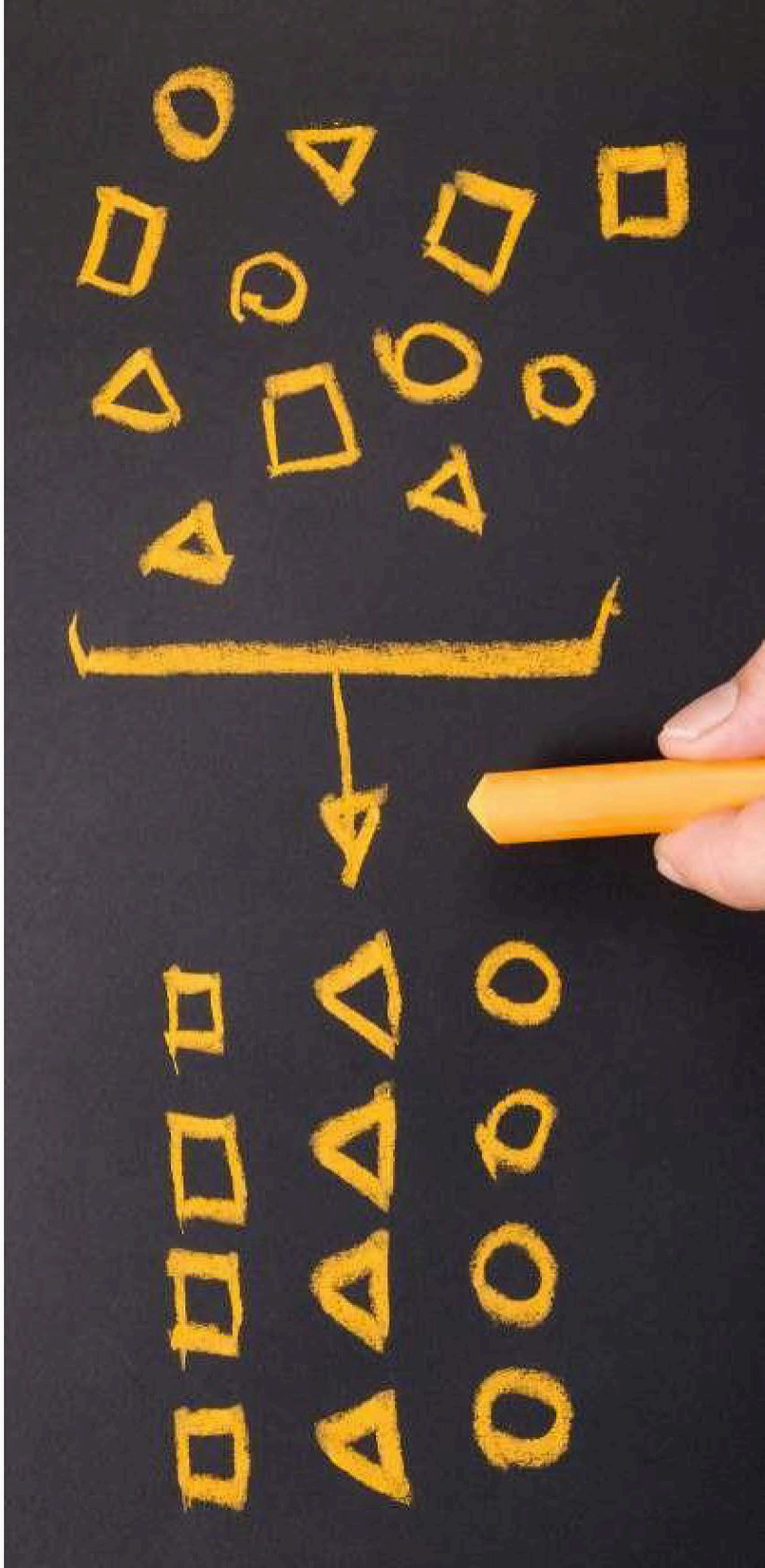


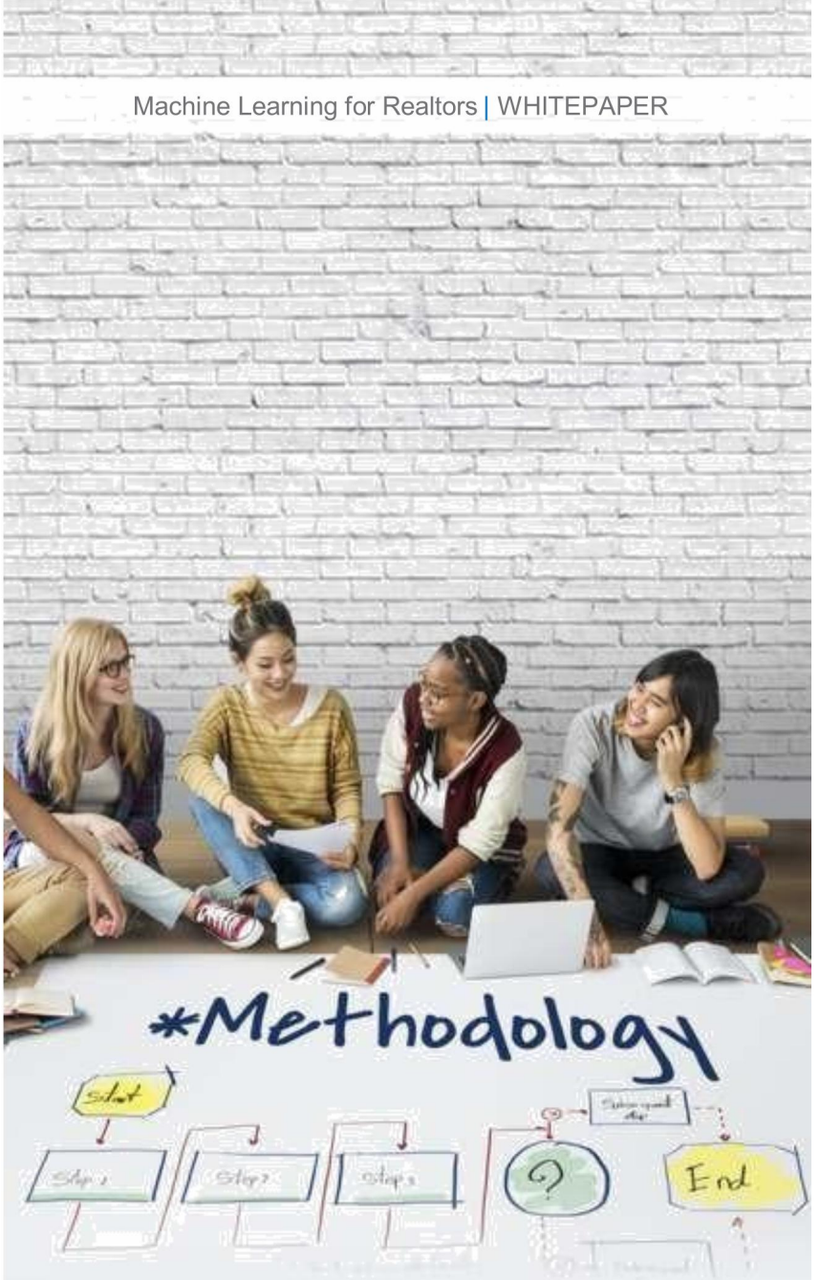
- **Embedded Method:** A combination of both filter and wrapper methods.

We chose to go with wrapper method as we had mixed set of continuous and categorical features. We used Backward Elimination process in wrapper methods. We eliminated few features using domain knowledge and applied machine-learning algorithms on the remaining features to do the filtering.

As feature selection was a one-time activity, time taken for training the model was not a concern. Accuracy of the model was an important factor we chose to consider. We used algorithms like Logistic regression; Random Forest and Support vector Machines and then chose features with high feature importance from each model. We then compared the top features obtained from each algorithm and then chose features from SVM as it gave good prediction results.






## Methods

The modelling pipeline is same for the three accelerators and the results varies only on the dataset filters and feature set filters.

**Propensity Prediction Model:** We used Support Vector Machines algorithm for customer's propensity prediction. We built different models for each line of business e.g. buy/sell/lease etc. The ROC-AUC is 0.73 for the SVM model with an accuracy of 0.90.

**Customer Behavior Analysis:** We initially chose K-Means Clustering algorithm to cluster customers with similar behavior. K-means is one of the simplest algorithm, which uses unsupervised learning method to classify a given data set into a certain number of clusters. It works really well with large datasets. The algorithm works iteratively to assign each data point to one of K clusters based on the features that are provided. Data points are clustered based on feature similarity.

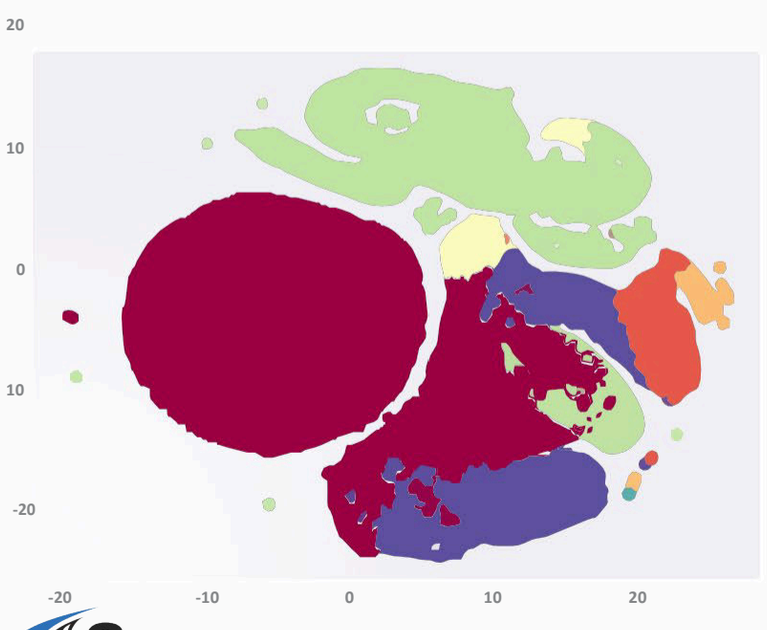
We built K-means clustering model on the data with best feature set from SVM model. We then used elbow method to obtain the optimal number of clusters (k).

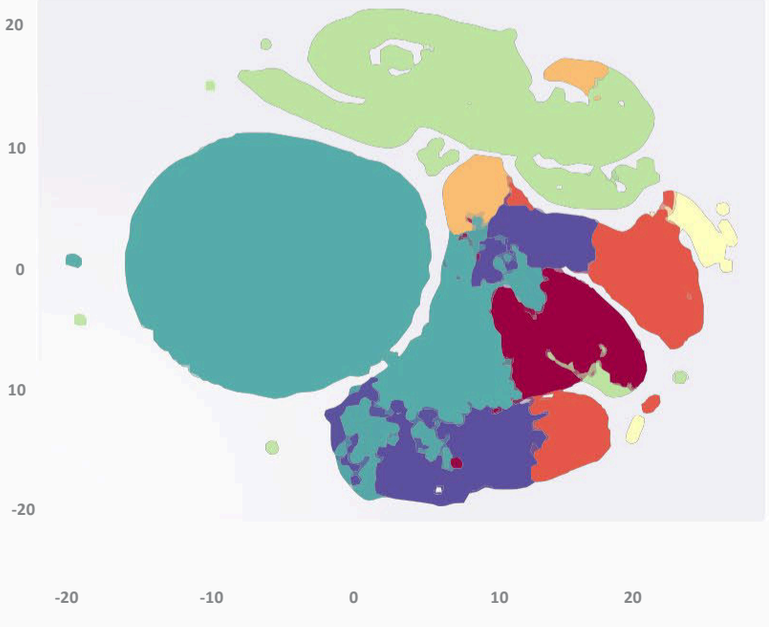
We then used Mini Batch K-means algorithm. Mini Batch K-means is far more useful in web applications where amount of data is huge and time available for clustering is very limited. Here the algorithm randomly chose small subsets of data called batches. Each point in the batch is assigned to a cluster and the cluster centers are updated. This is repeated for the entire data. Mini Batch K-means finds perfect balance between accuracy and computation time.

K-Means and Mini Batch K-Means Comparison: This comparison was done on a machine with Hexadecane core (16 core CPU) with 16 GB RAM. Using “Mini Batch K-means” instead of “K-means” algorithm for clustering improved the performance time by 97.5% whereas the accuracy did not vary much as evident from the t-SNE plots below.

T-SNE: t-Distributed Stochastic Neighbor embedding is a non-linear dimensionality reduction algorithm used for exploring high-dimensional data. It finds patterns in the data by identifying observed clusters based on similarity of data points with multiple features. The lower dimensional data has no identifiable features from original data and thus cannot be used to make inferences about them. It is thus only useful for visualization. In simple terms, t-SNE calculates similarity of data point x1 to data point x2 as a probability distribution p. Higher p values are nearer to each other compared to farther away points.

For k=7 clusters, below are the t-SNE plots for K-Means and Mini Batch K-Means respectively



## Silhouette score: 0.73

**Cluster 0** Propensity: 0.05

Low income (24K), not owner-occupied, large size dwelling the unit, younger in the sample. Mortgage is less than current home value. Buyer loyalty card users, Black -Hispanic, high Profitability, Sports Enthusiast less risk takers, low monthly payments, non-registrants politically, low retail trades, high external collection trades, security system owners

**Cluster 1** Propensity : 0.5372

Moderate Income range, more mortgage left to pay that current home value, multiple types of residential neighborhoods, Video gamers, Average monthly payments, worst status on last 6 months mortgage trade

**Cluster 2**

Has a child, Sports Enthusiast, Less educated in the sample, Debit card users, less travel time to work

**Cluster 3** Propensity : 0.1546

Income is above average, less sports enthusiasm, not a video gamer, dwelling size is less, least non-registrants politically in the sample, not a debit card users

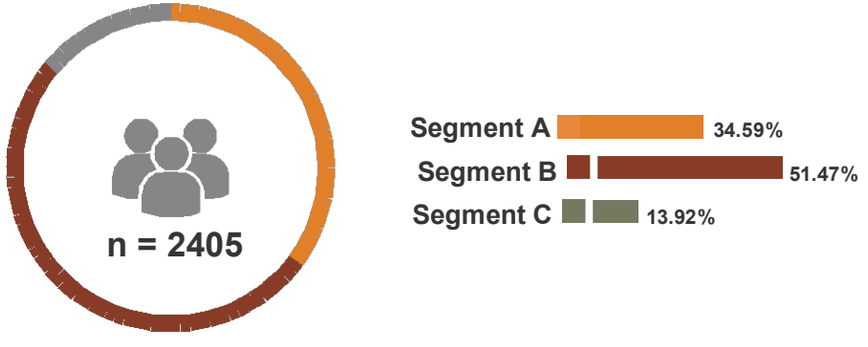
**Cluster 4** Propensity : 0.1836

Owner occupied, No child, High Income, Average age is high, lone to mortgage the Ratio is high, High-risk takers in the sample, highest monthly payments

**Cluster 5** Propensity : 0.03

Above average income, No Child, Low Profitability score, less white Hispanic, video Sports enthusiast, average monthly payments

## Population : Segmentation




| ParentID | Segment | Propensity To Buy - Score |
|---|---|---|
| 10000 | SEGMENTA | 0.04269638637 |
| 1001 | SEGMENTA | 0.225606838888 |
| 10026 | SEGMENTB | 0.0905146300794 |
| 1004 | SEGMENTB | 0.378420536349 |
| 1006 | SEGMENTB | 0.102577404602 |
| 10060 | SEGMENTB | 0.0170364907735 |
| 1009 | SEGMENTB | 0.158469454392 |
| 10099 | SEGMENTA | 0.213954500943 |
| 10106 | SEGMENTB | 0.0170364907735 |
| 10132 | SEGMENTA | 0.28183541231 |
| 10142 | SEGMENTB | 0.0170364907735 |
| 10147 | SEGMENTB | 0.0170364907735 |
| 10163 | SEGMENTB | 0.0817037201357 |
| 10168 | SEGMENTB | 0.289873987243 |
| 1017 | SEGMENTB | 0.0213456712509 |
| 10172 | SEGMENTB | 0.25841663216 |
| 1018 | SEGMENTA | 0.125323643576 |
| 10193 | SEGMENTA | 0.116530910092 |
| 10194 | SEGMENTA | 0.10411975253 |
| 10201 | SEGMENTB | 0.0170364907735 |
| 10210 | SEGMENTB | 0.0365949137123 |
| 10214 | SEGMENTB | 0.11346833656 |
| 10217 | SEGMENTB | 0.0170364907735 |
| 10222 | SEGMENTB | 0.0823612090007 |
| 10224 | SEGMENTB | 0.0170364907735 |
| 10225 | SEGMENTB | 0.0862374258782 |
| 1023 | SEGMENTB | 0.0243421051811 |
| 1024 | SEGMENTB | 0.338745921005 |
| 10248 | SEGMENTB | 0.0170364907735 |
| 10254 | SEGMENTB | 0.0586836559212 |
| 10257 | SEGMENTB | 0.0170364907735 |
| 10264 | SEGMENTB | 0.0170364907735 |
| 1027 | SEGMENTA | 0.373073702434 |
| 10285 | SEGMENTB | 0.0661340540995 |
| 1029 | SEGMENTB | 0.877500334905 |
| 10315 | SEGMENTB | 0.0170364907735 |
| 10318 | SEGMENTB | 0.355523698297 |
| 10323 | SEGMENTA | 0.25077163812 |
| 10324 | SEGMENTB | 0.00467580900011 |
| 10335 | SEGMENTB | 0.0170364907735 |
| 1035 | SEGMENTB | 0.00431741570616 |
| 10358 | SEGMENTB | 0.0190626918987 |
| 10360 | SEGMENTB | 0.0170364907735 |
| 10369 | SEGMENTB | 0.29716205426 |
| 10383 | SEGMENTA | 0.173420109482 |

**Segments**

| Segment A | Segment B | Segment C |
|---|---|---|
| Census: % Income $200-249k <=33.3% | Activities/Interests: Avid Runners Highly Likely | Avg # Credit Mortgage Inquiries (6 Months) Lower Avg # |

## Benefits

**Allows Personalized Communication:** Having a better understanding of their customers, the realtors can target each customer groups with a bespoke approach rather than a one size fits all technique. Targeting by specific group makes it easier to communicate with their customers with message relevant to them, providing a more personalized approach with appropriate marketing communications

**Enables Prospect Acquisition:** Having identified the profiles of the best and most profitable customers, realtors can then find look alike prospects and target them in an effective manner. These prospects will have a higher propensity to take up their offer and therefore provide a more cost-effective means of targeting new customers.

**Optimizes cost on Data resources:** Acquiring customer data from third parties can be an expensive procedure. Building an in-house ML-analytics solution also adds to the cost of the entire analytics procedure. Using a real-estate advisor ML analytics application can thereby optimize the cost involved.

**Provides Quick Analysis:** Real time reporting helps in building multiple reports using various subsets of features in a short time. It gives immediate access to the information sought, and bring new insights that allow Realtors to make decisions on what to do next.