Real Estate Market Analysis

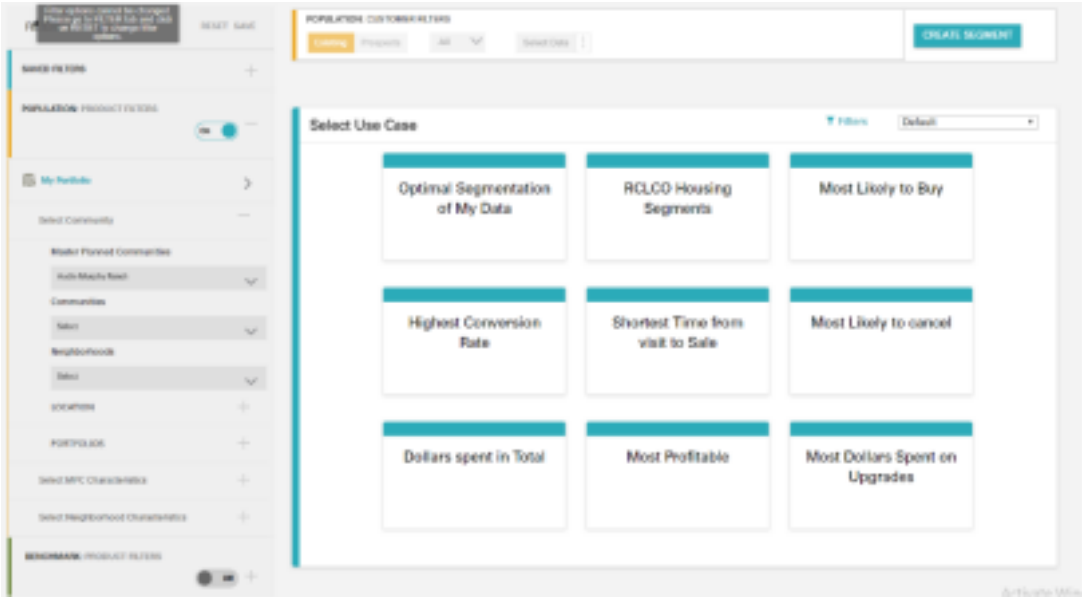
Use case:

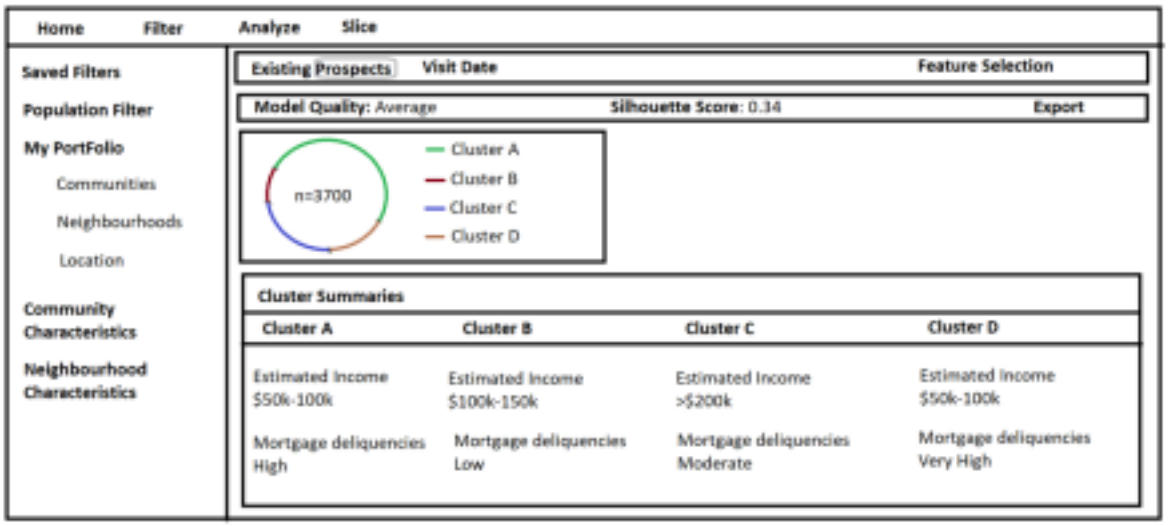
Realtors face challenges in finding right customers to sell or rent houses/lands/offices/buildings. They want to know the target customers and prospects who have higher propensity to take up their offers. They also want to group customers by shared characteristics such as demography, geography, psychography and behavior, so that they can develop different business strategies to engage with each customer groups depending on important characteristics/attributes of the group e.g. income, affluence, age, location and life stage.

Business Challenges:

- According to Norbert Winkeljohann, PwC Germany, the “sharing economy” - where people rent or borrow goods rather than buying or owning them - could generate revenues of US$335 billion by 2025. While youngsters are opting for services like Airbnb(short term lease), other end of the spectrum, old people need assisted living to keep themselves out of hospitals.
- As the consumers are less predictable now-a-days with traditional demographic personas, there is a need for Analysis tool which includes behavioral, psychographic, financial profiles as well
- Rapid urbanization is already having substantial economic, environmental and health effects on urban communities. Need for newer solutions like social housing etc
- With Increasing connectivity the need for office space is changing. Companies are operating completely in remote. This could have an impact on property prices in urban centers.
- We have to understand the customer's online behavior as new types of financial investors called FinTechs is transforming real estate. They are using AI technology to disrupt the financial industry. I.e., robo advisors are providing real estate advice for half the price of a broker
- The real estate market is increasingly impacted by government policies. In a globalized world, it is hard to anticipate how monetary, fiscal and planning policy, industry regulation and cross border legislation will affect markets. A better alignment of urban management and integrated planning is needed

Sample UI:





Implementation:

Having understood the challenges realtors face in marketing their offers, we designed and developed solutions which helped realtor in the following areas
1. Identify higher propensity customers and prospects
2. Analyze lead customers and prospects
3. Build communication strategies to approach customers and prospects

We developed three machine learning accelerators on 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 accuracy and used for final prediction.

We then designed three solutions to analyze different customers characteristics. We developed 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: Pre-built machine learning analytics helps realtors to quickly visualize and analyze different group's/segment's 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 line of businesses 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 to visualize and analyze different group's/segment's characteristics on the specific set of customer population they are interested in. This requires realtor to have depth understanding of domain to choose more relevant features for their line of business.

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

Enhanced self-services machine learning analytics: Enhanced self-service machine learning analytics helps realtors to visualize and analyze different group's/segment's characteristics on the specific type of customer population they are interested in. This service does not require realtor to have in depth domain knowledge to identify features for machine analytics as we already pre-identified features for each line of businesses. Important features are pre-identified for specific line of business e.g. buy or lease using domain knowledge and machine learning techniques.

The enhanced self-service machine learning provides control to realtor over features pre-identified for each line of businesses and type of customers that they want to use for machine learning analytics. Realtors can choose features from the available list of features. It also allows realtor to select subset of customers by applying various filters. The enhanced self-service analytics system then apply clustering techniques on the subset of data and the features selected by the realtor, and then present realtors with cluster analysis reports and propensity value for each customer to buy or lease.

1. **Data analysis** : We have *Client data* which has data records of realtor's customers and prospects based on their interaction with the realtors and *Customer data* which has all the financial, demographic, geo-demographic, psychographic and behavioral characteristics of all customers and prospects from a third party . We performed Exploratory Data Analysis on these datasets to understand the features and added new features wherever necessary. Below are few insights from Analysis:
   - Some features Like Year Built, Visit Month etc. have an impact on client to lease the property .
   - Customers visit more during summer and also probability of them leasing the property is also more.

- 13% clients are likely to lease , 34% of clients are likely to renew lease.

2. **Data Preprocessing**: Retained the relevant columns from client data after adding new columns based on the data analysis , Merged the Customer data with this Client data with ParentID as Key . Converted the data into usable format , removing the duplicates and missing value imputation .Reduced number of categories by binning the categorical features with too many categories and performed one hot encoding on all categorical features and filtered out some of numeric features using Pearson correlation.
3. **Feature Selection** : When there are too many features, we have to reduce the number of features since most of them may be noise which can affect the model performance. There are 3 different ways we can go about this.

   - *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 feedback loop. We recursively run the model with all the features and keep removing least important till we land with the best
     features or start with random set of features and keep adding new features till we get best model. Also, they use a classification performance of an classifier (like accuracy , ROC scores) to do the evaluation. Despite being computationally expensive , wrapper based are advantageous for giving better performances since they use the target in classifier for the feature selection .

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

Our Approach:
We chose to go with wrapper method as we have mixed set of features : continuous and categorical. We used Backward Elimination process of wrapper methods. Post data preprocessing we were left with dataset of around 180k records with 500 features. As feature selection is a one time activity, time taken for training the model is not really a concern. Accuracy of the model is an important factor we chose to consider. Logistic regression model is searching for a single linear decision boundary in your feature space, whereas a Random Forest is essentially partitioning your feature space into half-spaces using *axis-aligned* linear decision boundaries. Logistic regression is a classifier with a logarithmic cost function coupled with sigmoid as activation function. Optimization is easy for such a cost function as it is convex in shape. Only problem it has is overfitting. As the training dataset increases, it catches on the noise from the data to predict results. To minimize such a problem, we used L1 regularization which penalizes the cost function to reduce fitting problem. With a good fit, Logistic regression can work faster with accurate results. Kernel SVM, Random Forests(based on decision trees) are known to give more accurate results. We have to select the top features and thus interpret the results in terms of feature importance. Biggest advantages of using Decision Trees and Random Forests is the ease in which we can see what features or variables contribute to the classification. Kernel SVM transforms the data points from input space to kernel space. Instead of getting the data points in feature space, we get the dot products of data points with support vectors using only kernel functions, which is then used to classify the data point. This property

makes SVM suitable for high dimensional data. Our current case has 500 dimensions(high). In Kernel space, coefficients of hyperplane or feature weights have no meaning and thus feature importance will be unknown if we choose to proceed with it. So, using kernel SVM is not suitable for our case. But there is a version of SVM called SVC(support vector classifier) which is based on liblinear library optimized for linear kernels. In SVC, coefficients of the hyperplane represent the feature weights. We used this version in hope of an accurate model and getting feature importance . Based on initial analysis on data and research, Random Forest Classifier and Linear classifiers like Logistic Regression, SVC can be better binary classifiers.

Random forests classifier, an ensemble learning method for classification that operate by constructing multiple decision trees and merges them together to get a more accurate and stable prediction. Also measures the relative importance of each feature on the prediction which can be good to use while considering feature selection to remove unwanted noise.

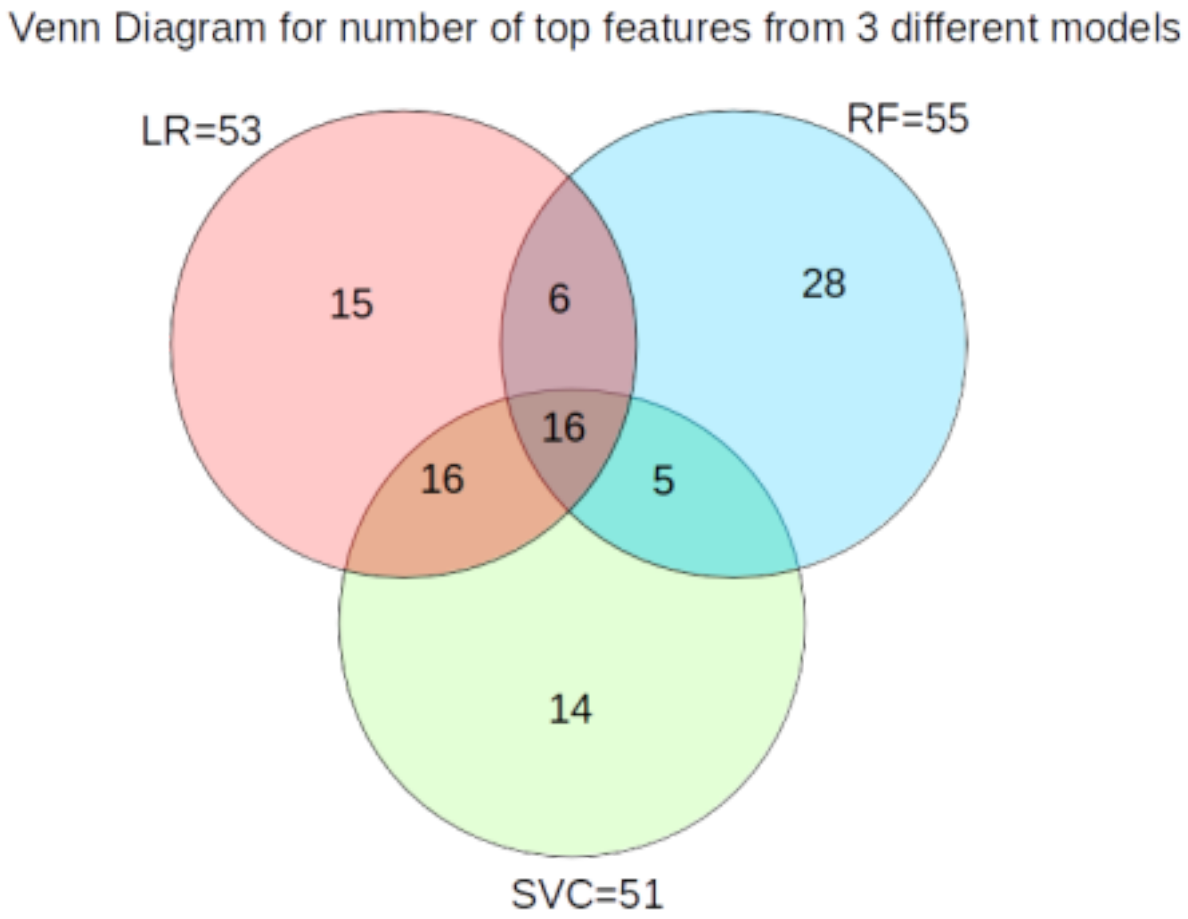
Logistic Regression , one of the most popular machine learning algorithms for binary classification, a simple linear model ,when used along with lasso(L1) regularization ,it tends to produce sparse solutions by forcing unimportant coefficients to be zero.

Support Vector Classifier: Linear SVM is a linear classifier which is characterized by the normal to the hyper-plane dividing positive and negative instances. Components of the normal with higher absolute values have a larger impact on data classification.

Based on these feature importance, feature weights from random forest classifier and linear models , we will be able to choose our feature set.

4. **Modeling and Clustering** :

Before we train our models with each of chosen algorithms, in order to get the best set of features, tuned the hyper parameters in each of the models using Grid search CV which performs Exhaustive search over specified parameter values for an estimator. After we obtain data with our best set of features, to group customers based on their similarities we use Clustering. Comparing the features from Logistic Regression, Random Forest, Support Vector Classifier:



K-Means Clustering: K-means is one of the simplest algorithm which uses unsupervised learning method to classify a given data set through 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.

Built K-means clustering model on data with best feature set from each model, used elbow method with silhouette scores for k=2 to 15 to obtain the optimal number of clusters(k).

Mini Batch K-means: 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 we 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.
Comparision:

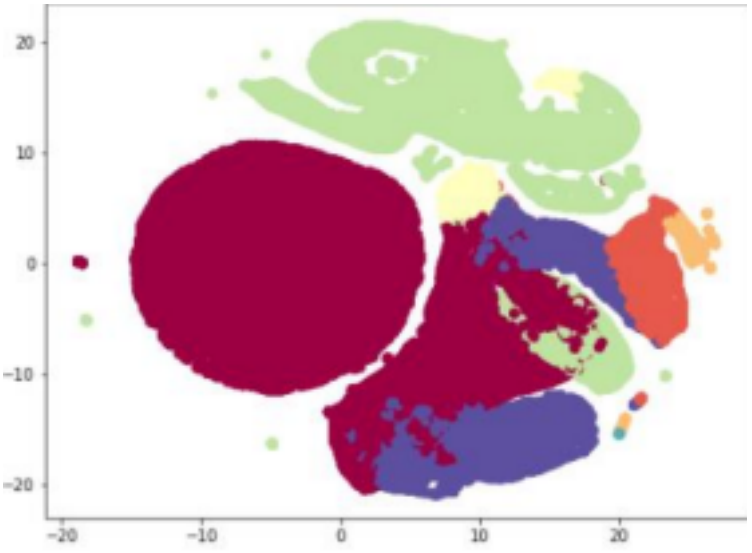
Comparision is done on a machine with Hexadeca core(16 core CPU) with 16Gb RAM. Using “Mini batch k-means” instead of “k-means” algorithm for clustering improved the performance time by 97.5% where as the accuracy did not vary much as evident from the t-SNE plots below.

T-SNE : t-Distributed Stochastic Neighbor Embedding is a non-linear dimensionality

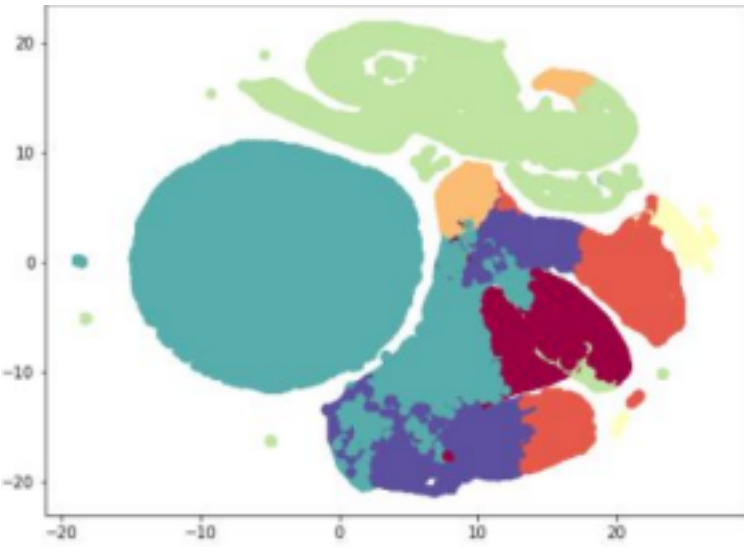
reduction algorithm used for exploring high-dimensional data. It finds patterns in the data by identifying observed clusters based on similarity of data points with multiple features. The lower dimensional data has no identifiable features from original data and thus cannot be used to make inferences 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 kmeans and MiniBatch kmeans



respectively



Post EDA: We then performed Compared the features, AUC, Accuracy, Precision from each of the model. Clustering after SVC gave max silhouette score of 0.78.
Benefits:

1. ***Personalize 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

2. ***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.

3. ***Optimizes cost on Database resources*** : Credit Statistics based companies like Experian, Transunion CIBIL, Equifax collect data pertaining to the financial, demographic, geo demographic, psychographic, behavioral characteristics for an individual household. Collection of data is done by credit reports, census, surveys etc. This data can be purchased at bulk on zip code level. Let's say we brought the data from Experian. We need customer data from the realtor so that we can map it to the customer profiles from Experian for Analysis. As the realtor clients increase, we could do Analysis on realtor data, his competitors data and also Entire Experian Household data. Entire Experian data is expensive for a single realtor. Also, realtors should be able to analyze the data each time it is updated .

4. ***Quick Analysis***: Real Time Analysis for the prospective clients with say highest income range, business with whom will be most profitable, most likely to buy, most likely to lease, most likely to renew lease contract, who has most spends recently etc is provided. Realtor data can be updated regularly with visit dates from clients, new clients, Lead source details etc. As the computations are done in real time, we get the most latest results

Results:

1. Dynamic Clustering for Self-Service Analytics: Realtors can effectively select the data using filters like location, community, neighborhood to perform clustering on the filtered data. Clustering is performed on each individual cluster again to identify behavior patterns. They can use the results from important attributes to understand the behavior of the customers in a specific cluster. Time we spend online for the analysis to complete must be as minimum as possible. We have achieve very high gain in computation times using mini batch kmeans without losing much of accuracy for clustering.

2. Pre-built Machine Learning Models: These can be used for quick analysis for a specific use case. We identify most important features for a use case and perform clustering on data with important features and propensity scores. Using SVC as the best model gave an accuracy of 90.7 percent.