Look-alike Modeling

Finding Similar Customers Through AI

PerceptiveAnalytics

Marketing Analytics | Tableau Consulting | Excel Apps
New York • Dallas • San Francisco • Hyderabad
Perceptive-Analytics.com
(646) 583 0001 cs@perceptive-analytics.com

Look-alike Modeling – A Case Study

Executive Summary

- Business & marketing professionals are in a constant hunt for the potential prospects and they invest time, money and effort in filtering to arrive at the potential prospects with high chances of conversion. This reduces cost of acquisition of customers and increases profitablity.

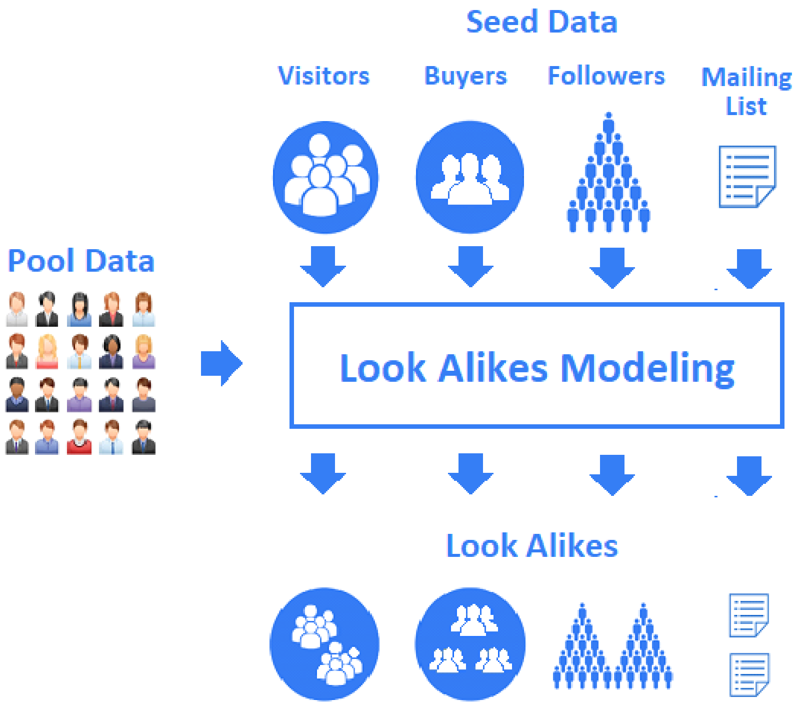
- Look-alike modeling is an effective targeting system that reduces your investments required to acquire new customers and helps your business to expand with a stable business model.

- In this casestudy, we present a machine learning based lookalike model for a bank to target prospects for investing with the bank. This was done based on previous campaign data. We gave details of sequence of steps required for building the model like analysis of various data imbalance handling techniques, modelling methodology, validation and interpretation of results. The final model pipeline has performed well over the existing targeting system in use giving a 450% improvement.

What is a Look-Alike?

Look-alike audiences are the prospects who are having similar traits, behavior like your already existing customers.

Look-alike modeling is a process that essentially helps you in finding look-alike audiences of your best, most profitable customers. It is a modelling approach that can be used by marketers to define customers who are most likely to engage with their marketing messages or activities. This model analyzes and considers common behaviors or traits among the current customers and seeks potential customers who have similar characteristics.



Definitions

Seed Data is data of existing customers based on whom we want to find look-alike audiences.

Pool Data is the customer database, in which we would look for customers who are look-alikes to seed data. Pool data could be collected from various sources.

Extended Look-Alikes Audience are the look-alike audience generated by the model from pool audience, based on seed data.

Benefits of Look-Alike Modeling

Look-Alike model plays a key role in making business and marketing related decisions. Helps in understanding existing customer base and expand business reach by only focusing on your best customers with a stable business model.

Effective Targeting

Look-Alike model helps businesses and marketers to execute better marketing campaigns by limiting their focus to those prospects who are similar to the target customers on whom the business is interested in.

Lower Acquisition Cost

Customer Cost Acquisition (CAC) cost is, in general, approximately 6-7 times costlier than Customer Retention Cost (CRC). But by relying on look-alike modeling, businesses can reduce CAC as they would only spend their marketing efforts on potential customers (look-alikes) who are more likely to convert.

Loyal and Profitable Customer Base

Look-alike modeling helps you in building a highly profitable customer base, by allowing you to target the look-alikes of those customers with high Customer Lifetime Value (CLV) ensuring a highly profitable customer base for your business, in the long run.

Data

This Look-alike model is built on internal marketing campaign data of a bank. The bank had conducted a marketing campaign on a fraction of its client base to find out which of them are likely to invest with the firm. The campaign-related data(seed data) and data of the clients who were not part of the campaign(pool data) were taken into consideration.

The data consists of various client, social, economic, demographic and campaign-related attributes. The pool and seed data contains 32K and 10K client's data respectively. The seed data is imbalanced with only 11% of clients who were campaigned had actually invested.

Objective

To acquire customers more effectively with less marketing spend and effort.3

As the campaign data is imbalanced (11% conversion), the objective here is to:

- Propose a solution to handle the imbalanced nature of the campaign data.
- Build a look-alike model based on this campaign data and find the look-alike audience of the potential investors of the campaign from the remaining client base.
- Increase the conversion rate of the client investments by reducing the marketing spend, effort and time with look-alike audience generated.

Challenge faced with imbalanced dataset

Handling the imbalanced campaign data before building a model, was one of the key challenges we came across. When dealing with imbalanced data, a predictive model developed using conventional machine learning algorithms like Decision Trees, Logistic Regression, etc. could be biased and inaccurate.

These algorithms improve accuracy by reducing error without considering the class distribution in data. They tend to predominantly predict the majority class data correctly while the features of the minority class are treated as noise and are often ignored.

Because of the class imbalance in input data, there is a high probability for misclassifying the minority class (churn), compared to majority class (not churned)

Solutions to Deal with Imbalanced Dataset

We dealt with all these following approaches to handle imbalance in the dataset and picked the one that fits our data best.

Algorithmic Approach

These approaches are based on modifying the machine learning algorithms that decide the look-alikes, so they can handle the imbalanced datasets with better performance.

Ensembling

This improves the performance of classification by constructing several models on original data that have diverse learning techniques from each other and then aggregating their predictions. The diversity in model's learning can be possible by building each base model on a different dataset, relating to the same classification.

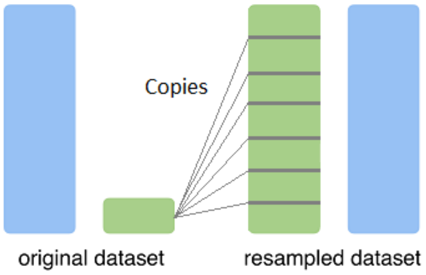
Custom Cost Function Techniques

In imbalanced datasets since the algorithms are biased towards majority classes, we can compensate it by the adjusting varying the penalty for miss-classification in major and minor classes.

Data Approach

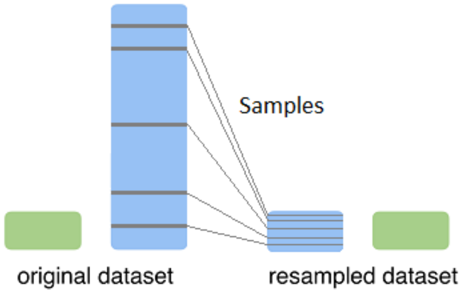
These approaches are based on re-sampling the data before it is given as input to the algorithms, in order to mitigate the effect caused by class imbalance.

Naive Oversampling



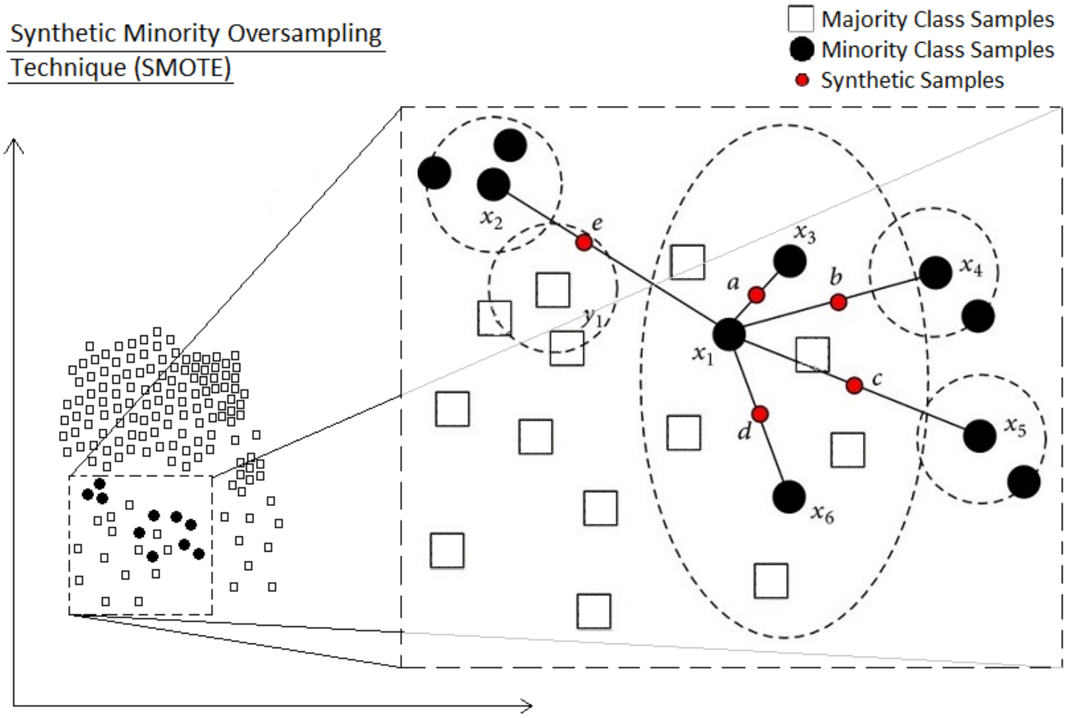
Duplicating the minor class observations, to make them equal to the number of major class observations. This increases the likelihood of overfitting since it replicates the minority class observations.

Naive Undersampling



Aims to balance class distribution by randomly eliminating majority class examples until majority and minority class instances are balanced out. In the process, it discards a lot of potentially useful information, making it difficult for the models to learn the actual patterns in data.

Synthetic Minority Oversampling Technique (SMOTE)

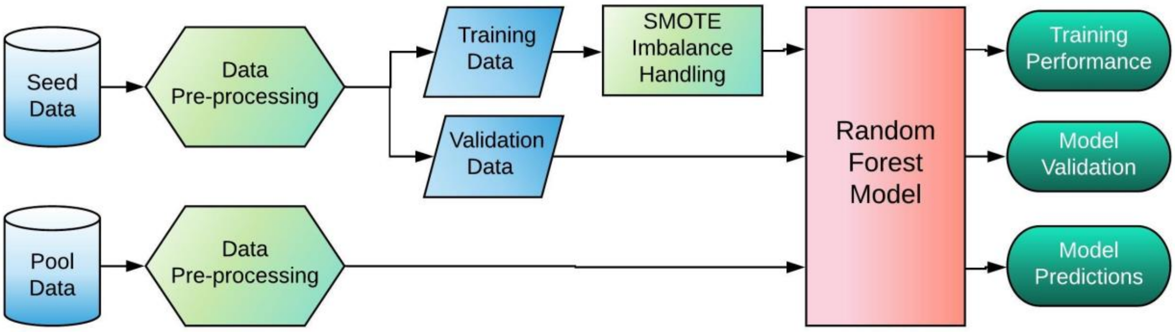


SMOTE (Synthetic Minority Oversampling TEchnique)

Generates synthetic data using a k-nearest neighbors classifier rule on the original samples, thus overcoming the drawback of overfitting in Naive Oversampling. SMOTE implementation, while generating synthetic data, will not make any distinction between the samples based on whether they are easy or hard to classify.

Modelling Methodology and Results

Seed data will be used to train & validate the model. Predictions would be made on pool data. We have built a Random Forest Machine Learning model on the Pre-processed and sampled train data. There are two main steps that are involved before the modelling.



Data Pre-Processing

In the data preprocessing step the seed data was analyzed and handled for missing, outlier data. Required feature engineering has been done on the seed data, all the categorical ordinal and nominal categorical variables are label and one-hot encoded respectively.

Data Imbalance Handling

The train data is imbalanced in its class distribution, and it needs to be handled before the model was built. Since we only had one dataset to work on, if we do an ensemble, there cannot be diversity between models, and hence the ensemble model would most likely not ensure better performance than the base classifier models. Of all the data imbalance techniques, sampling techniques are most appropriate for this dataset given the nature of data. SMOTE algorithmic sampling of generating synthetic data performed best when compared to all other sampling techniques.

For imbalanced dataset, the normal classification metrics like accuracy and ROC-AUC fails to correctly assess the model performance, therefore we also evaluated the model using additional parameters like F1-Score and PR-AUC which assess the model performance correctly even in the cases of imbalanced data. We built and compared performances of advanced machine learning algorithms like Neural Network, Random Forest, Light GBM, SVM, Logistic Regression.

Random Forest model built on SMOTE sampled data performed better than any other combination of sampling and machine learning algorithms

Using the trained model, we predicted which customers among those in the test data are most likely to invest their money with the bank. Very high validation **PR-AUC score (91%)** and validation **F1-Score (83%)** in below table imply that model we built is performing really well in making predictions on unseen data.

Metric	Performance Scores	Validation Scores
Accuracy	0.88	0.84
F1-Score	0.87	0.83
ROCAUC	0.93	0.92
PRAUC	0.98	0.91
Precision	0.95	0.9

Using previous targeting method only 11.11% of the campaigned clients had actually invested and the rest 88.89% are only contributing to addional marketing spend without any return, ie..in order to have ‘x’ number of conversions, we needed to reach ‘9x’ potential customers. But, using look-alikes targeting method, we only need to reach ‘2x’ potential customers, thus reducing additional marketing spend, time and effort.

Effectiveness of previous targeting system = Conversion/Reach = 1/9 = 11.11%

Effectiveness of Look-Alike targeting system = Conversion/Reach = 1/2 = 50%

Look-Alike targeting system is 450% is effective than previous targeting system!

Conclusion

Look-alike modelling is really a powerful tool that makes targetting more effective and precise. It empowers the marketers to spend their budget on the right target customers.

For every campaign the seed data gets updated with data of converted customers and model will be retrained, thus increasing the effectiveness of the model further helping optimize and save huge sums of marketing spend, time and effort by focusing on right audience.

Thereby reducing Cost to Acquire a Customer (CAC) significantly lower than expected profitability from a customer in his entire lifetime (CLV), producing rapid growth in business expansion activities with higher stability in business model.

About Perceptive Analytics

Perceptive Analytics is a Data Analytics Company recognized as a Top 10 Emerging Analytics Company. It is the winner, Fidelity Data Challenge in which 54 analytics companies participated. It also received an award at Netflix Hackathon at Tableau Conference, 2018. The clients we served include Morgan Stanley, Johnson & Johnson, Amex, Wells Fargo, and PepsiCo to name a few. We work with clients as their analytics department or alongside internal teams to deliver long-term competitive advantage.

Winner, Fidelity Investments Data Challenge

Top 10 Emerging Analytics Companies to Watch

Advanced Analytics

- Customer Analytics
- Marketing mix modeling
- Attribution modeling
- Customer segmentation
- Response modeling
- Churn modeling
- Next basket analysis

Business Intelligence & Reporting

- Visualizations & Dashboards
- ETL (Extract, Transform and Load)
- Data warehousing
- Cloud computing (AWS, Azure, Google Cloud)

Schedule Free Consulting Call

Contact

Chaitanya Sagar, CEO
cs@perceptive-analytics.com
+1 (646) 583 0001

Marketing Analytics | Tableau Consulting | Excel Apps
New York • Dallas • San Francisco • Hyderabad
Perceptive-Analytics.com