# Predictive Ensemble Model for Dataset Selection to Accurately Predict Movers

Rounak Mehta, Ankur Manikandan, Aman Agarwal, Peizhen Gu
New Opportunities in Bigdata

## Introduction

In the United States, on an average, only about 7% of homeowners move annually. This is a real problem for the realtors because they have thousands of contacts in their networks, but can reach out only to a few hundred a year. Due to the current system, realtors miss out on potential customers.

First a Durham, NC based real estate tech startup founded in 2014 is fundamentally changing the way realtors target their next customer. Their key objective is to predict when and why people will move.

In order to make such accurate predictions, First uses extensive datasets – Epsilon and Acxiom

## Problem Statement

The problem required the construction of a predictive model to assist in the determination of which of the two provided datasets: Acxiom or Epsilon is superior.

Determining the superior dataset will help First gain better insights into their predictions. Also, not only will their prediction accuracy increase, but it will also help them to optimize their expenses on data acquisitions, i.e. get the best return on investment.

The problem required building an imbalanced classifier to predict the movers. In the Epsilon and Acxiom datasets 2.65% and 2.64% comprised only of movers.

## Methods

**Data Wrangling and Exploratory Data Analysis**

The operations that were performed on either of the datasets are as follows:

1. The Acxiom dataset was imputed using the median. The Epsilon dataset was provided with the missing values imputed by First.
2. For both the datasets, some of the categorical variables were collapsed.

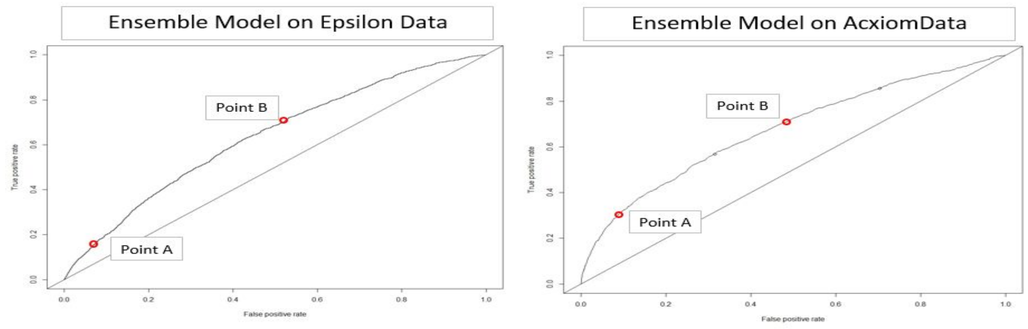
**Feature Engineering and Selection Algorithm**

For the Acxiom dataset, we had feature engineered the area available per person as it proved to be an important predictor. This feature was already present in the Epsilon dataset.

The variable selection was performed by creating an average ranked list of variables in their order of importance after performing a Random Forest algorithm and determining the individual AUC.

| Ensemble Model using scores from Logistic Regression and Random Forest | | | | Metrics | Epsilon | Acxiom |
|---|---|---|---|---|---|---|
| Point A: Classifying top 10% as Positive | | | | | | |
| | | Test Classification | | True Positive Rate | 17.29% | 31.53% |
| | | 5685 | 49001 | Positive Predictive Value | 4.40% | 8.30% |
| Condition | 1446 | 250 | 1196 | Accuracy | 87.87% | 89.03% |
| | 53240 | 5435 | 47805 | Odds Ratio | 1.80 | 4.15 |
| Point B: At a True Positive Rate of 70% | | | | | | |
| | | Test Classification | | True Positive Rate | 69.99% | 70.44% |
| | | 28702 | 25984 | Positive Predictive Value | 3.53% | 3.85% |
| Condition | 1446 | 1012 | 434 | Accuracy | 48.57% | 52.95% |
| | 53240 | 27690 | 25550 | Odds Ratio | 2.11 | 2.57 |
| Area Under the Curve (AUC) | | | | | **0.635** | **0.695** |

| Ensemble Model using scores from Logistic Regression and Random Forest | | | |
|---|---|---|---|
| Point A: Classifying top 10% as Positive | | | |
| | | Test Classification | |
| | | 4626 | 41649 |
| Condition | 1218 | 384 | 834 |
| | 45057 | 4242 | 40815 |
| Point B: At a True Positive Rate of 70% | | | |
| | | Test Classification | |
| | | 22269 | 24006 |
| Condition | 1218 | 858 | 360 |
| | 45057 | 21411 | 23646 |




## Predictive Classification Model

The model used to determine the best of the two datasets was an ensemble model. The model was a combination of a 10 fold cross-validation logistic regression algorithm without regularization and a random forest algorithm.

The model was trained using 60% of each of the two datasets and it was trained using the remaining, i.e. 40% of each of the two datasets. . To maximize the performance of the overall model, the majority class (non-movers) was down sampled to a 1:1 ratio. This was performed primarily due to the random forest algorithm. The random forest algorithm over fits the original dataset. Hence, we had to change the ratio of movers to non-movers of the training set to prevent overfitting.

The ensemble model was tested on unseen data where the ratio of movers to non-movers was maintained the same as the two datasets.

## Conclusions

After running the ensemble model, we produced the confusion matrix and the ROC curve for the corresponding datasets.

With an AUC value of 0.695 that we obtained from the Acxiom dataset, it is clear that the Acxiom dataset outperforms the Epsilon dataset.

Therefore, we make the recommendation to First that they must use the Acxiom dataset to increase their prediction accuracy and to optimize their expenses on data acquisition.

## Acknowledgement

We would like to thank Prof. Daniel Egger and Mr. Andrew Born from First for giving us the opportunity and support to execute this project. We would also like to extend our thanks to the rest of the First team.