Maximizing Profit through Direct Mail

Presented by George Wilson
KPMG Super Day 10/4/2018

Adapted from 5W Strategists Case Competition
Collaborators: Sam Beadles, Charles Chen, Andrew Gerin, and Tykai Martin

Agenda

1. Context & Objectives
2. Executive Summary
3. Data Walkthrough
4. Modeling Approach
5. Results and Recommendations

Who is the target customer for direct mail?

Our client uses direct mail to acquire U.S. customers. They have 40,000 results from a previous campaign, and want to know which **5,000 individuals to target next**.

Objectives

Maximize profit from mailing campaign using previous knowledge about respondents

Out-scoping

- Are **other mediums** than direct mail more effective?
- Is there an **optimal number** of packages to send to maximize profit?

Approach

We developed and tested several models to rank the most profitable individuals on a unknown test set

Our analysis assumes:

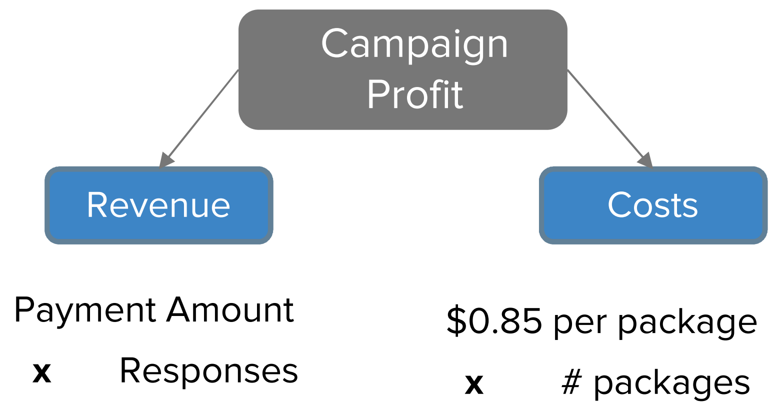
- Original and testing population a representative samples of a larger population
- No effect of history on the experiment

Executive summary

Direct mail is a profitable source of new customers.

Given an response rate of 12%, the biggest driver in revenue will be increasing total responses.

Of several models considered, LightGBM was submitted to the competition. The model increased client profit by 51%, and lifted response rate by 39.9%



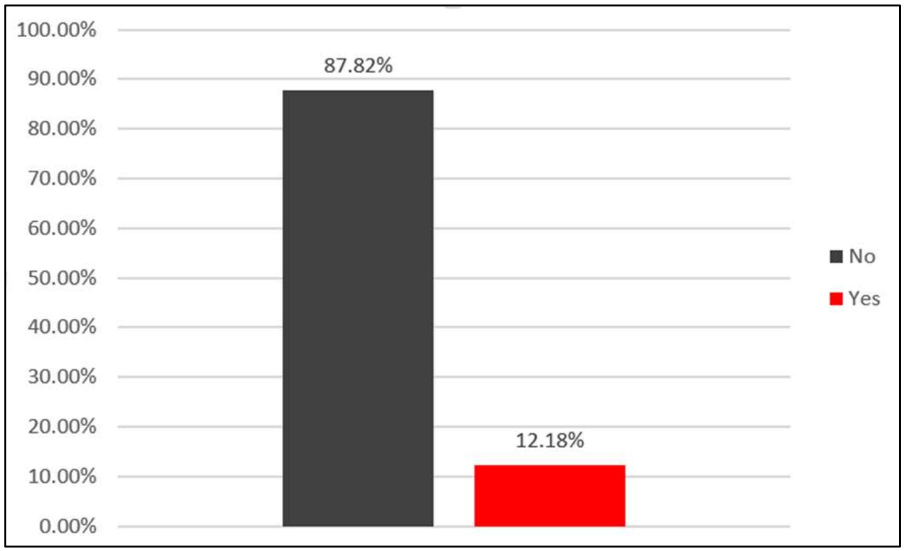
Method	Response Rate	Responses (per 5000)	Profit	Response Rate Lift
Baseline	12.18%	609	$73,700	N/A
LightGBM	17.1%*	855*	$111,800*	39.93%*

*Based on competition results on unseen data

Maximization of revenue depends strongly on responses

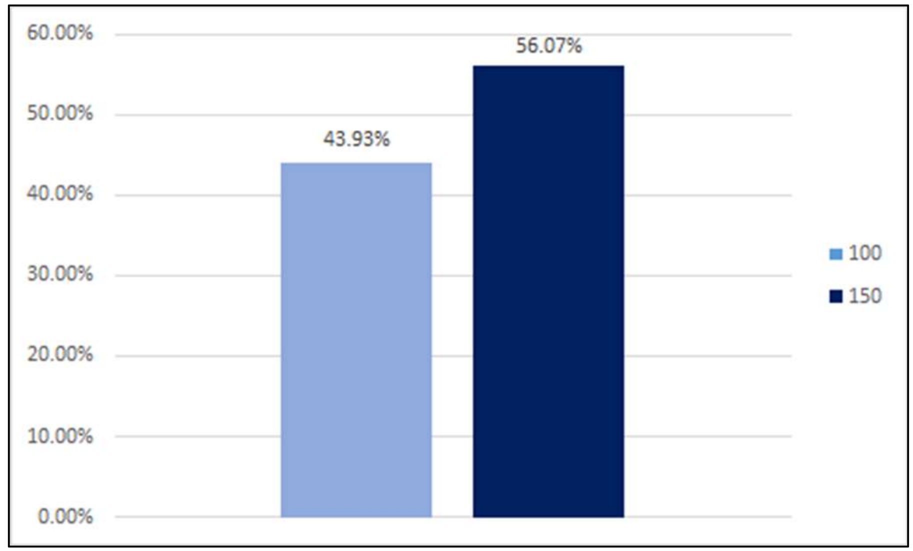
Campaign Profitability: $14.52 in average profit per package mailed

Response Rate



40,000 mailings resulted in **4,872 responses**

Payment Amount ($)



Expected value of response = $128

Data cleaning

Ambiguous Variables

- No data dictionary
  - Percent_Professional encoded a single value
  - Motorcycle_ind 95% zeros
  - Zero variance between some variables
- 189 features may result in noisy dataset

Miscoded Data

- Explored summary statistics
- Identified -1 coded instead of N/A
- Recoded indicators to Boolean
- Created dummy variables for factors to be ingested

N/A Values

- No all customer information is readily available
- Removed columns with greater than 2,000 missing values
- Performed mean imputation

Feature selection using random forest does not identify many payers

- Two initial Random Forest Classifiers
  1. All features with dummy variables
     - Random noise most important feature
  2. Using only quantitative features

	Importance
WCr_Avg_Median_Home_Value	0.051644
WP_county_pop_density_census	0.040057
Income_Pred_Narrow_Mid_Pts	0.036626
W5r_Median_Income	0.033876
W5r_Curr_Home_Value	0.032235
Random	0.031030

- Random Forest with top five features by feature importance

5-Fold CV Results:

- 87.2% Accuracy
- 50.5% - 52.5% AUC
- **1% Recall**

Takeaway: Algorithm performs well based on accuracy; however, it fails to differentiate true positives from false negatives.

Advanced modeling: what response rate can be expected?

LightGBM Model: fast, gradient boosted model
Good for: Large datasets and imbalanced classification problems
Limitations: Difficult to explain, slow parameter tuning

Original Objective: Identify top 12.5% of test set

- On a holdout set of 8,000 individuals
    - Top 12.5% responded **19.3% of the time**
    - 5-Fold Cross Validation: 16.05% - 23.9%

Implication of Model:

Method	Expected Response Rate	Responses (per 5000)	Profit (All $100 Payments)	Profit (All $150 Payments)
Worst CV LightGBM	16%	800	$75,000	$115,000
Avg. CV LightGBM	19.3%	965	$92,000	$140,000
Best CV LightGBM	23.9%	1,195	$115,000	$179,000

Recommendations & competition results

- Direct mail acquisitions offer a profitable new customer base
    - Selecting particular individuals may increase revenue given resource constraints
- LightGBM submission increased response rate by **39.9%** on unseen data
- Final results fall within expected profit projected by cross-validated predictions

Total Profit: $116,050

Baseline Profit: $73,700

Value Added: $42,350

Predictive modeling in this context performs well, but has room for improvement through parameter tuning and enhanced data cleaning

Questions?

Appendix

Future Work

Due to time constraints, the following were considered but not implemented:

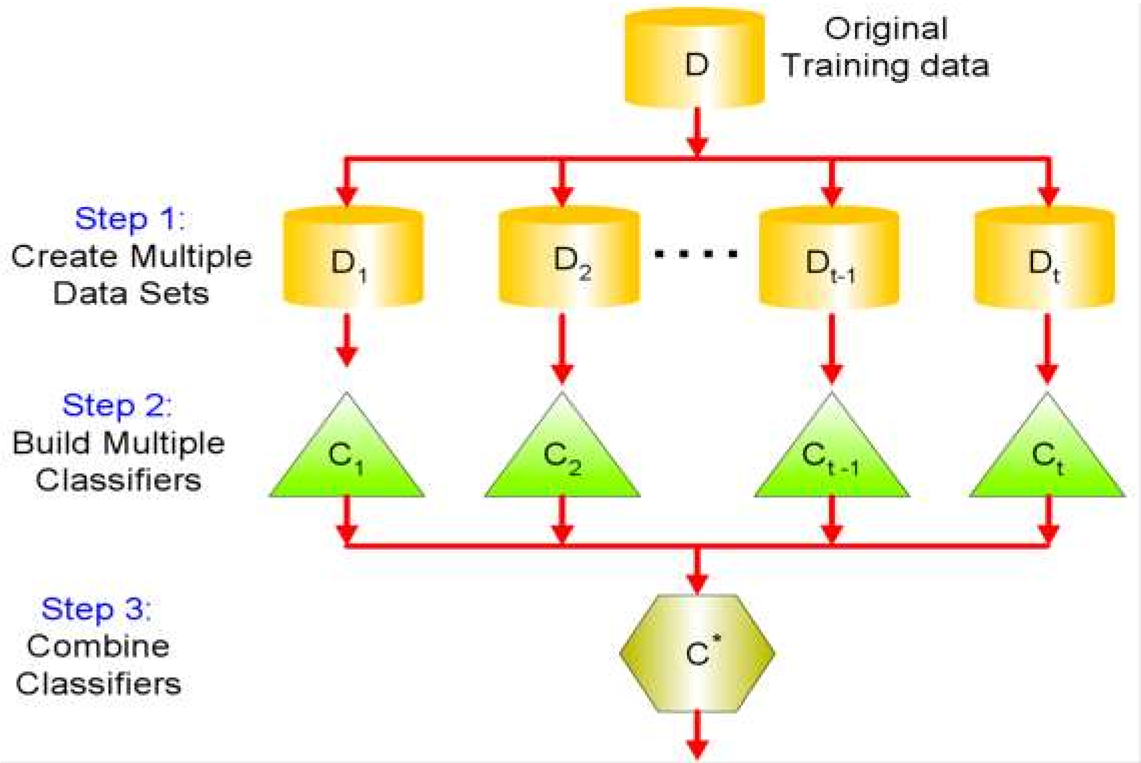
- Impute using multiple imputation methods (i.e. MICE)
- Explore different **feature selection techniques**
    - PCA for quantitative variables, high correlation filter, Smooth Lasso
- Engineer new features
- Integrate model with expected value of payment
    - See logistic regression analysis
- Tune model parameters
    - Grid Search

Limitations & Lessons Learned

- Advanced algorithms are difficult to explain, “black-box” effect
- Ambiguous features and dirty data plagued modeling success
  - It is difficult to create a model without fully understanding the data and its attributes
- More data exploration and visualizations are necessary
- Managing response rate vs. expected value of the response
  - Given low response rate, classifying responses is more important than higher payments

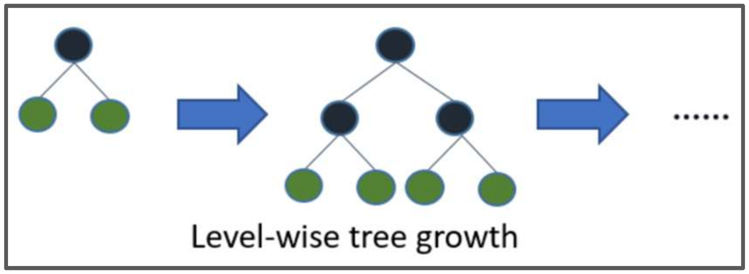
Random Forrest Classifier

- **Ensembled Method**: averages over diverse decision trees
- Each tree is based on bootstrapped sample
- Each node split is based on P random queries

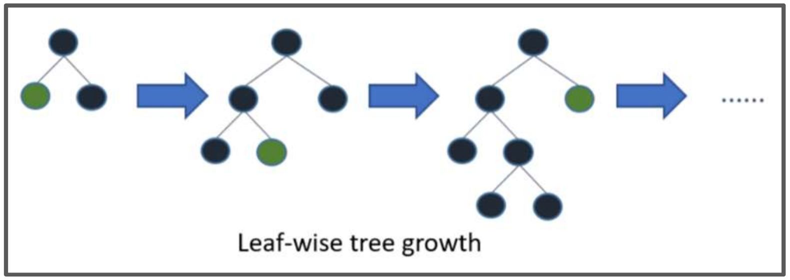


Light GBM Model

1. Gradient Descent
   - Moving towards the point of minimum error
   - Problem: how large of increments ..too fast then overshoot, too slow then very long model...adjusted via learning rate
2. Boosted
   - Start with weak learner. Improve by weighting misclassified attributes heavier
3. Leaf wise growth increases accuracy



Traditional Boosting Algorithms



Light GBM

Light GBM Prediction Submission

count	40000.000000
mean	0.186273
std	0.044859
min	0.125947
25%	0.156580
50%	0.168773
75%	0.223307
max	0.318032

Predicting payment amount using marriage indicator increases expected value of mailings

We can use whether or not someone is married to predict how much they pay:

- **Observation**: Campaign Association Flag (CAF) tells us the size of payment.
- **Observation**: (CAF) and indicator of marriage have strongest association
- **Action**: Impute Campaign Association Flag using married indicator.
  - HHLD_Married_flg had data for all rows

We generate two probabilities for married vs not married:

Probability of CAF being 1 when married indicator is 1: **0.69**

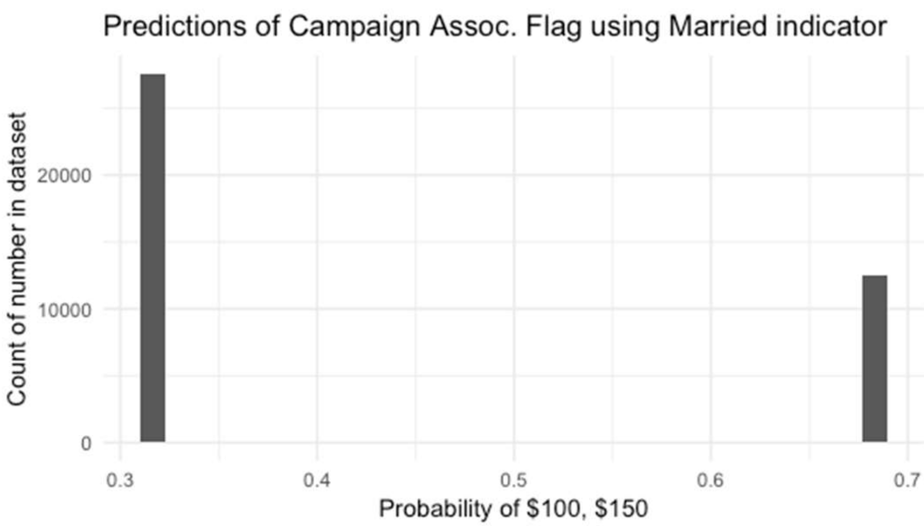
Probability of CAF being 1 when married indicator is 0: **0.31**

This allows us to distinguish between expected values of responders

- EV(CAF=1) = $135
- EV(CAF=0) = $115

Payment amount pay depends on married indicator

Plot of proportion of married vs not married, and relevant probabilities



Logistic regression results

```
Call:
glm(formula = recoded_pay$Campaign_Assoc_Flg ~ recoded_pay$HHLD_Married_flg,
    family = "binomial", data = recoded_pay)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5261  -0.8798   0.8649   0.8649   1.5077

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -0.74964    0.05212  -14.38   <2e-16 ***
recoded_pay$HHLD_Married_flgY  1.54011    0.06466   23.82   <2e-16 ***
```