Title

## presented by

Team Member 1

Team Member 2

Team Member 3

1. Header…

The purpose of this study was to create a model that predicts a person’s probability of making a purchase. Ideally, this model will help us decide on who and where to focus our marketing.

The most influential variable in correctly predicting a person’s purchase decision was \_\_x\_\_. Specifically, people tended to make a purchase when their value on \_\_x\_\_ was high. This makes sense because …

Here is a graph of \_x\_\_ over the time which the data were collected:

[placeholder graph]

I have eliminated some of the outlier data points because it appears those points were anomalies or errors and really do not belong to the population being studied. In fact, there were a number of negative values for the \_\_\_y\_\_ variable, which is of course impossible. I tested 3 separate models for the rest of the data: binning the \_\_z\_\_\_ (with bins of equal size, bins were defined as 0-1, 1.1-2, 2.1-3, 3.1-4), treating \_\_z\_\_ as a binary (with a cutoff of 3), and treating \_\_\_z\_\_\_ as a continuous variable.

One interesting observation is that \_\_\_z\_\_\_ tends to increase with time. Therefore, the effect of \_\_\_\_z\_\_ on the purchase decision is confounded with time. This could be due to changing economic conditions, increased brand recognition of our company, or the government regulation that went into affect in June of last year. Below is a summary of the models’ performance.

|  |  |  |  |
| --- | --- | --- | --- |
| \_\_\_z\_\_\_ | Overall Accuracy | Recall  (what % of those who made a purchase were correctly identified by the model) | Precision  (what % of purchase predictions were correct) |
| Contiuous | .61 | .52 | .06 |
| Binned | .70 | .55 | .27 |
| Binary | .63 | .43 | .70 |

Based on these results, we moved forward with the model that uses binned values of \_\_\_z\_\_.

Furthermore, regardless of \_\_\_\_z\_\_\_, income was helpful in predicting the probability of making a purchase. In particular, those less than 30,000 annual household income (i.e. >55) are less likely to make a purchase.

1. Recommendations

* When \_\_\_z\_\_\_ is high, any degree of marketing seems ineffective.
* When \_\_\_z\_\_\_ is mid-range, the efficacy of marketing increases. The following guidance should be followed when deciding who to call:
  + Direct mail marketing to households greater than 4 (median probability of purchase increase by .35 over those smaller household).
    - Start with largest households and move to smaller
    - In addition, the mail should be sent with an offer for zero down, zero interest for 6 months. Probably of purchase drops sharply as the window of zero interest decreases as shown by this graph.

[PLACEHOLDER GRAPH]

* When \_\_\_Z\_\_\_ is low, the efficacy of marketing increases substantially and mailings should be pursued at a much higher pace. Specifically who to focus on follows the same guidance as the mid-range values for \_\_z\_.
  + In addition, those with more facebook friends are more likely to make a purchase. The plot illustrates this, where the red line is predicted probability of a purchase, and each point represents an individual.

[PLACEHOLDER GRAPH]

1. Methods

Three separate models were created based on how \_\_z\_\_\_ was treated. Not all available variables were included in the model because …

A random forest model was used in each case, with a max-depth of 5 layers to prevent over fitting and to keep things simple. Various values for this parameter were attempted, but somewhere near 5 seemed to hit highest recall levels without overfitting. No appreciable difference was noticed for values of 4 or 6 when using 5-fold cross-validation. Over sampling was used on the training dataset in order to not bias the model towards a “no” prediction. Simple decision trees were also attempted but consistently performed worse (5-20% less recall of “yes, purchase made”).