DETERMINING IF MARKETING LEADS WILL CONVERT

PAUL BUTLER

THE PROBLEM

- Marketing team at X Education needs a strategy for improving lead conversion rate.
- Lead Conversion is currently at 38%.
- Marketing team wants a system that can rank, and sort leads into hot or cold categories based on the likelihood of conversion.
- Once this has been achieved new strategies can be developed with less uncertainty.

DATA SET

- Data for 9240 prospects who were in contact with the marketing team.
- 37 Variables including:
  - Subjective employee ratings of each prospect
  - Web activity of prospects
  - Lead source and marketing channel information
  - Basic location info
  - Permissions for follow-up contact

FEW OBVIOUS RELATIONSHIPS APPARENT IN THE DATA

- Subjective ratings were only predictive for extreme values.
- The only major employment category more likely to convert than not were working professionals.
- No apparent major difference between conversion in India, than outside of India.
- No major difference between conversions by city.
- Only one modest correlation in web activity.

FEW OBVIOUS RELATIONSHIPS CONT.

Subjective Lead Quality	Total Leads	Leads Converted	Percent Converted
Data Missing	4767	1024	21.48%
Worst	601	12	2.00%
Low in Relevance	583	477	81.81%
Not Sure	1092	266	24.36%
Might Be	1560	1179	75.58%
High in Relevance	627	603	96.17%

FEW OBVIOUS RELATIONSHIPS CONT.

Occupation	Total Leads	Leads Converted	Percent Converted
Data Missing	2690	370	13.75%
Unemployed	5600	2441	43.59%
Working Professional	706	647	91.64%
Businessman	8	5	62.50%
Housewife	10	10	100.00%
Student	210	78	37.14%
Other	16	10	62.5%

FEW OBVIOUS RELATIONSHIPS CONT.

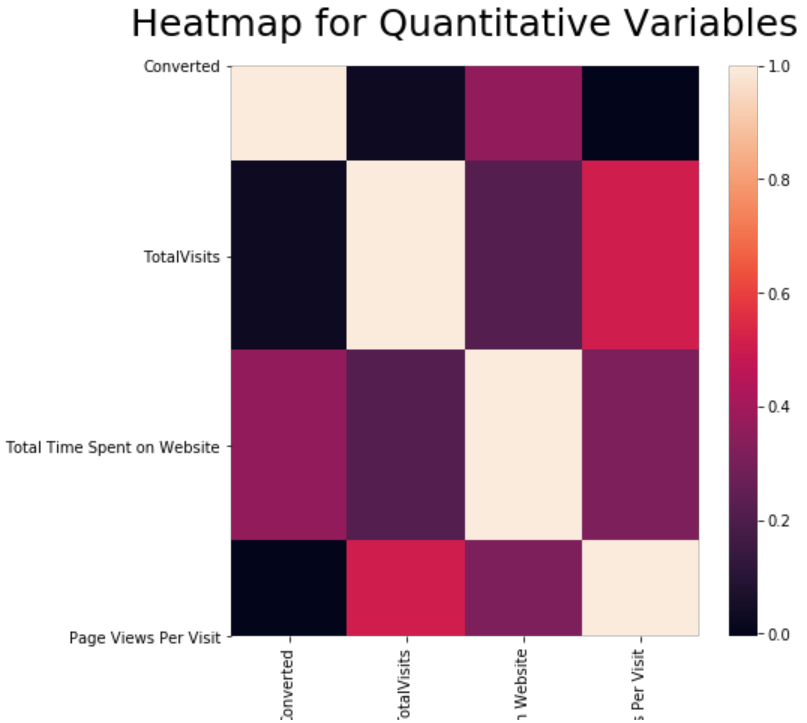
Country	Total Leads	Leads Converted	Percent Converted
India	6492	2401	36.98%
Other	2748	1160	42.21%

FEW OBVIOUS RELATIONSHIPS CONT.

City	Total Leads	Leads Converted	Percent Converted
Data Missing	3669	1257	34.26%
Mumbai	3222	1309	40.62%
Thane & Outskirts	752	338	44.94%
Other Cities	686	276	40.23%
Other Cities of Maharashtra	457	201	43.98%
Other Metro Cities	380	155	40.79%
Tier II Cities	74	25	33.78%

FEW OBVIOUS RELATIONSHIPS CONT.

- Only continuous data was website activity
- Total Visits has modest correlation with Page Views Per Visit.
- No other obvious correlations



FINDING THE NONOBVIOUS RELATIONSHIPS

- Use of Machine Learning – Logistic Regression
- Model overview:
  - Binary Classifier of 0 for no conversion and 1 for conversion.
  - Also report the predicted chance in order to rank leads.
- Technologies used
  - Stats model for model evaluation.
  - Scikit-learn for model building.

MODEL TECHNICAL DETAILS

- Split data using stratified 80/20 train/test split.
  - This allows part of the data to be unseen by the model until testing – this allows us to confirm accuracy of model.
- Recursive Feature Elimination –Allows us to select the most useful features.
- Manual elimination of nonsignificant features using stats model –remove features that are not contributing well to model.
- Check Variable Inflation factor on remaining factors – ensure model is not inaccurately based on a small number of features.

MODEL ACCURACY

- Model is 90% accurate.
- This means that 90% of predictions on new data should be classified correctly.
- Other evaluation metrics used:
  - Cross validation – ensuring repeated trials of the model perform accurately using different sets of data.
  - Confusion matrix – ensuring model does not have significant differences between predicting different types of outcomes.

FINAL REMAINING FEATURES

- Final Model Contained 11 Features
  - Features are ranked from most predictive to least predictive
- 7 Positive Predictors (Green)
  - Presence of these features more predictive of a lead converting
- 4 Negative Predictors (Red)
  - Presence of these features more predictive of a lead not converting

Feature Rank	Feature	coef	std err	z
1	Tags_Closed by Horizzon	5.0557	1.024	4.939
2	Tags_Lost to EINS	4.4834	0.847	5.296
3	Tags_Will revert after reading the email	4.1441	0.390	10.639
4	Tags_Interested in other courses	-3.9045	1.030	-3.793
5	Tags_Ringing	-3.7259	0.428	-8.714
6	Tags_Already a student	-3.2659	1.060	-3.082
7	What is your current occupation_Working Professional	2.8423	0.938	3.030
8	Lead Quality_Worst	-2.3113	0.721	-3.204
9	Last Notable Activity_SMS Sent	2.1505	0.219	9.819
10	Asymmetrique Activity Score_15.0	1.5885	0.255	6.240
11	What is your current occupation_Unemployed	1.4446	0.195	7.419

SAMPLE OUTCOME FORMAT

- The model outcome format includes four pieces of information

1. Lead number of prospect
2. Lead Score: a 0-100 rating of the chance for lead conversion
3. The predicted outcome for the model
    1. 0 = predicts no conversion
    2. 1 = predicts conversion
4. The actual outcome
    1. 0 = Lead did not convert
    2. 1 = Lead converted

Lead Number	Lead Score	Predicted Outcome	Actual Outcome
624430	99	**1**	**1**
598020	98	**1**	**1**
657686	97	**1**	**1**
651611	97	**1**	**1**
595073	93	**1**	**1**
585273	70	**1**	**0**
613102	39	**0**	**0**
593831	25	**0**	**0**
641725	9	**0**	**0**
586770	1	**0**	**0**

CONCLUSIONS

- The conversion/non-conversion of leads can be successfully predicted from a small number of features.
- Leads can be ranked by the chance of conversion.
- Machine learning can be used to predict lead conversions better than other relationships in the data.