EMPLOYING CLASSIFICATION MODELS TO IDENTIFY POTENTIAL CUSTOMERS FOR A CARAVAN INSURANCE POLICY

VIVEK APPADURAI

FOUNDATIONS OF DATA SCIENCE WORKSHOP

SPRINGBOARD.COM

CARAVANS

- Called Trailers or Mobile Homes in the US
- Very popular in Netherlands
- Usually hitched onto a car for camping vacation trips

DIRECT MAIL OR POSTAL ADs

- 56.9% of total mail volume in the United States is postal advertising
- $46 billion spent by US businesses on Direct Mail campaigns in 2014
- Average household receives 19 direct mail catalogues per week
- 42% of recipients read postal advertisements
- 14.1% of individuals aged 45-54 respond <- key demographic

Junk Mail

- 44% of Junk Mail discarded without being opened or read
- 4 million tons of wasted paper per year – 32% recovered for recycling
- 33% of Americans find Direct Mailing intrusive

SOLUTION

Leveraging Customer Information and Machine Learning techniques to make better predictions on potential customers

DATASET

- Generated by the Dutch Data Mining company, Sentient Machine Research
- Used in the COIL 2000 challenge organized by Computational Intelligence and Learning Cluster in the year 2000
- Submitted to the UCI Machine Learning Repository
- Real World Dataset with demographic and Socioeconomic data from 5,822 (Training Set) & 4,000 (Test Set) customers
- Multivariate: 85 fields + Prediction

Available Data

DEMOGRPAHIC INFORMATION (Based on Neighborhood Zip code)

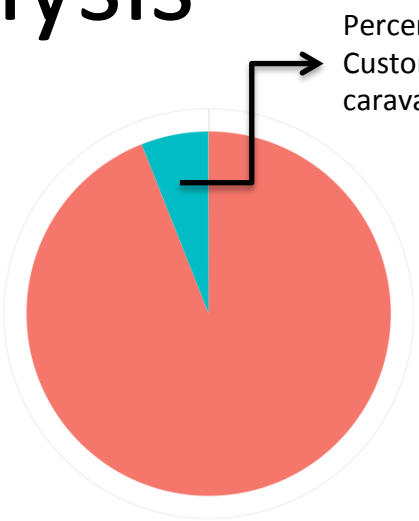
- 1 MOSTYPE Customer Subtype see L0
- 2 MAANTHUI Number of houses 1 ? 10
- 3 MGEMOMV Avg size household 1 ? 6
- 4 MGEMLEEF Avg age see L1
- 5 MOSHOOFD Customer main type see L2
- 6 MGODRK Roman catholic see L3
- 7 MGODPR Protestant ...
- 8 MGODOV Other religion
- 9 MGODGE No religion
- 10 MRELGE Married
- 11 MRELSA Living together
- 12 MRELOV Other relation
- 13 MFALLEEN Singles
- 14 MFGEKIND Household without children
- 15 MFWEKIND Household with children
- 16 MOPLHOOG High level education
- 17 MOPLMIDD Medium level education
- 18 MOPLLAAG Lower level education
- 19 MBERHOOG High status
- 20 MBERZELF Entrepreneur
- 21 MBERBOER Farmer
- 22 MBERMIDD Middle management
- 23 MBERARBG Skilled labourers
- 24 MBERARBO Unskilled labourers
- 25 MSKA Social class A
- 26 MSKB1 Social class B1
- 27 MSKB2 Social class B2
- 28 MSKC Social class C
- 29 MSKD Social class D
- 30 MHHUUR Rented house
- 31 MHKOOP Home owners
- 32 MAUT1 1 car
- 33 MAUT2 2 cars
- 34 MAUT0 No car
- 35 MZFONDS National Health Service
- 36 MZPART Private health insurance
- 37 MINKM30 Income < 30.000
- 38 MINK3045 Income 30-45.000
- 39 MINK4575 Income 45-75.000
- 40 MINK7512 Income 75-122.000
- 41 MINK123M Income >123.000
- 42 MINKGEM Average income
- 43 MKOOPKLA Purchasing power class

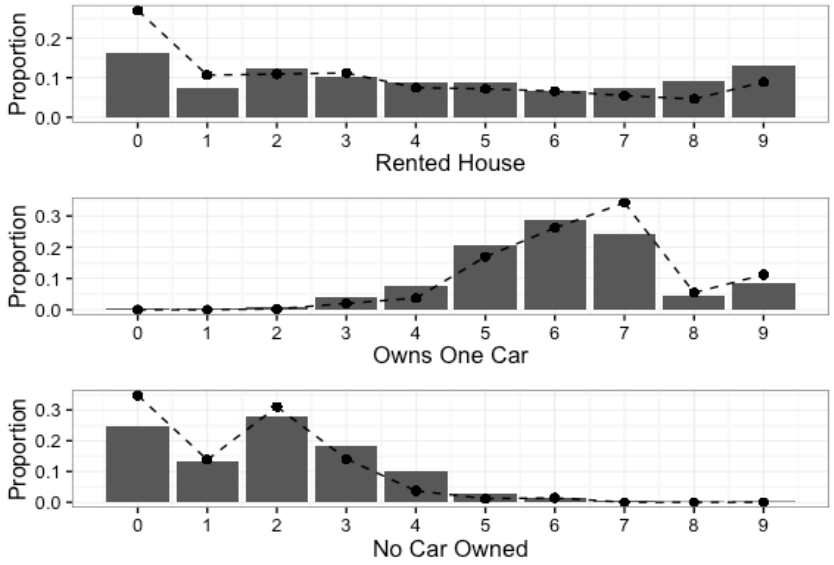
PURCHASE HISTORY (Unique to each Customer)

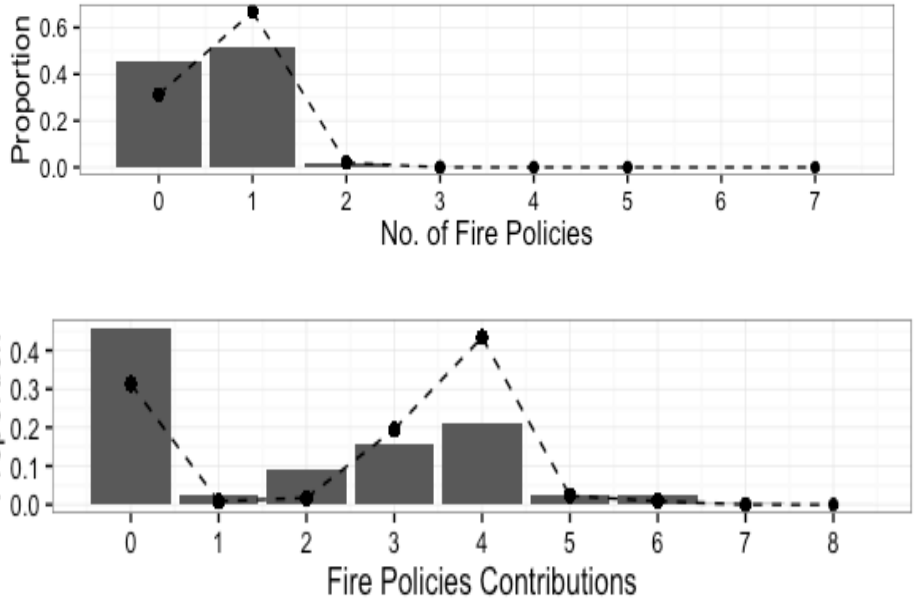
- 44 PWAPART Contribution private third party insurance see L4
- 45 PWABEDR Contribution third party insurance (firms) ...
- 46 PWALAND Contribution third party insurance (agriculture)
- 47 PPERSAUT Contribution car policies
- 48 PBESAUT Contribution delivery van policies
- 49 PMOTSCO Contribution motorcycle/scooter policies
- 50 PVRAAUT Contribution lorry policies
- 51 PAANHANG Contribution trailer policies
- 52 PTRACTOR Contribution tractor policies
- 53 PWERKT Contribution agricultural machines policies
- 54 PBROM Contribution moped policies
- 55 PLEVEN Contribution life insurances
- 56 PPERSONG Contribution private accident insurance policies
- 57 PGEZONG Contribution family accidents insurance policies
- 58 PWAOREG Contribution disability insurance policies
- 59 PBRAND Contribution fire policies
- 60 PZEILPL Contribution surfboard policies
- 61 PPLEZIER Contribution boat policies
- 62 PFIETS Contribution bicycle policies
- 63 PINBOED Contribution property insurance policies
- 64 PBYSTAND Contribution social security insurance policies
- 65 AWAPART Number of private third party insurance 1 - 12
- 66 AWABEDR Number of third party insurance (firms) ...
- 67 AWALAND Number of third party insurance (agriculture)
- 68 APERSAUT Number of car policies
- 69 ABESAUT Number of delivery van policies
- 70 AMOTSCO Number of motorcycle/scooter policies
- 71 AVRAAUT Number of lorry policies
- 72 AAANHANG Number of trailer policies
- 73 ATRACTOR Number of tractor policies
- 74 AWERKT Number of agricultural machines policies
- 75 ABROM Number of moped policies
- 76 ALEVEN Number of life insurances
- 77 APERSONG Number of private accident insurance policies
- 78 AGEZONG Number of family accidents insurance policies
- 79 AWAOREG Number of disability insurance policies
- 80 ABRAND Number of fire policies
- 81 AZEILPL Number of surfboard policies
- 82 APLEZIER Number of boat policies
- 83 AFIETS Number of bicycle policies
- 84 AINBOED Number of property insurance policies
- 85 ABYSTAND Number of social security insurance policies
- 86 CARAVAN Number of mobile home policies - **Target Variable (1/0)**

Exploratory Analysis

- Small number of policy owners in the population
- Best to check for proportion of policy owners at each variable level
- Intuitive sense suggests to look for customers who are married, own a home and car, fire policy, third party health insurance and makes high contribution to each policy
- Avoid Farmers, Laborers and low level educated people





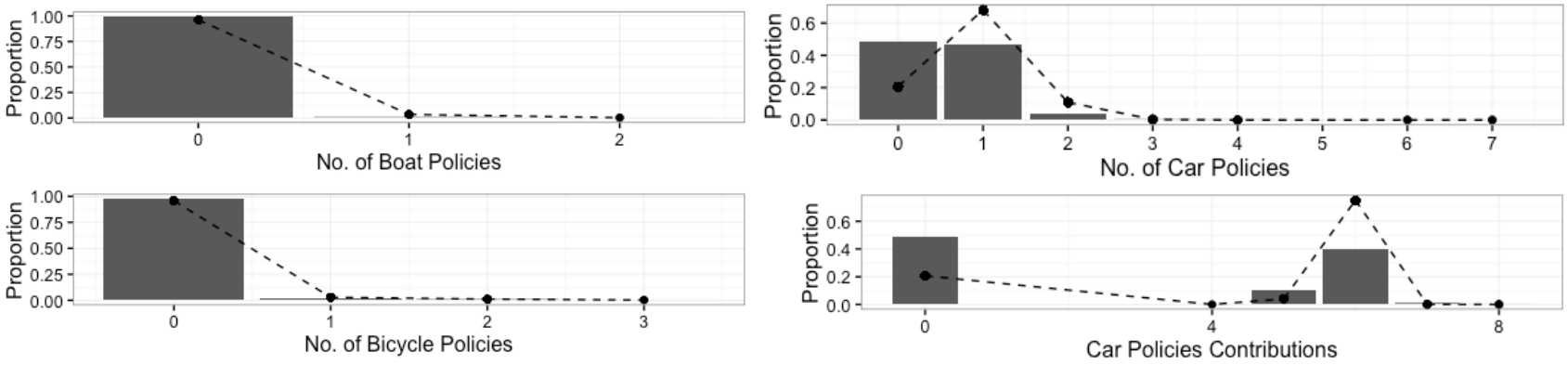


Classification Methods Employed

- Logistic Regression with Forward and Backward variable selection
- Logistic Regression with Random Forest variable selection
- Naïve Bayes Classifier
- Support vector Machine, Linear and Radial Kernels
- Random Forest Classifier
- Decision Tree Classifier

Importance of Automatic Variable Selection Algorithms

- Automatic Variable selection mostly confirms our initial intuitive expectations but show predictive power from unexpected variables which make sense in hindsight



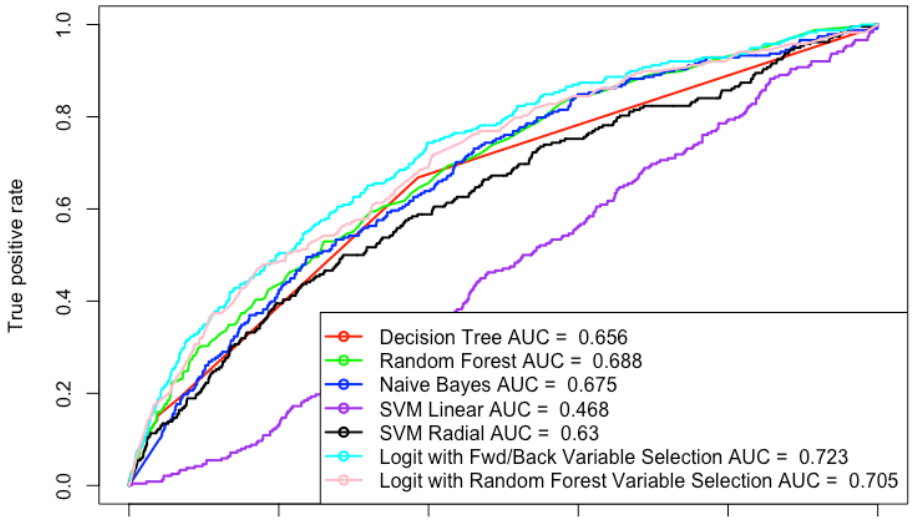
- People who own boats and boat policies do not show a higher proportion of policy owners but they score highly in automated variable selection
- Number of car policies and contributions is correlated, someone with more policies and small contributions to each will look similar to someone with a single policy but a high contribution towards it but the latter has a higher purchasing power

Performance Comparison

- Logistic Regression with forward and backward variable selection gives highest area under curve, Logistic Regression with Random Forest Variable selection is a close second
- We chose a 15% false positive rate to make 664 predictions out of 4,000
- Identify 104/238 potential customers correctly at a 44% accuracy

Confusion Matrix

	FALSE	**TRUE**
0	3,202	560
1	134	104



Conclusion

- We utilized the power of exploratory analysis of multivariate datasets, automated variable selection and machine learning for a real world marketing problem
- We compare different models and choose the best predictive algorithm
- We instruct the marketing associates in charge of direct mail advertising to target married customers with a home, third party health insurance, high purchasing power and previous indulgence in leisurely activities
- We suggest them to avoid low education level neighborhoods, farmers and people who insure farming equipment, singles and people living in rental homes or no evidence of car ownership

References

- Direct Marketing Association http://thedma.org
- USPS Household Diary Study, 2014
- United States Environmental Protection Agency https://www3.epa.gov
- Pew Internet and American Life Project
- www.StuffDutchPeopleLike.com
- Github Project Page: https://github.com/vaqm2/SpringBoard/tree/master/CapstoneProject
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+(COIL+2000)
- COIL 2000 Challenge Page: http://liacs.leidenuniv.nl/%7Eputtenpwhvander/library/cc2000/report2.html
- Hadley Wickham's dplyr and ggplot2

1. https://cran.r-project.org/web/packages/dplyr/index.html
2. http://ggplot2.org/