Deloitte.

# **Insurance Analytics** - analýza dat a prediktivní modelování v pojišťovnictví

Pavel Kříž

Seminář z aktuárských věd
MFF
4. dubna 2014

# Summary

1. Application areas of Insurance Analytics
2. Insurance Analytics process
3. Modelling techniques (CRT in more detail)

---

4. Case study – Identification of customers for X-selling campaign
    - Business and data understanding
    - Predictive modelling
    - Cost-benefit analysis

# Areas of application in insurance

# Main application areas of Insurance Analytics

| | |
|---|---|
| Cross-selling, Up-selling | • Which customers are most likely to buy more? => classification analysis |
| Retention | • Who is most likely to leave? => classification analysis |
| Fraud detection | • What claims are likely to be fraudulent? => cluster analysis |
| Claim segmentation | • Which claims are likely to become large? => regression analysis, cluster analysis |

# Main application areas of Insurance Analytics

**Pricing**

- What are expected claims/risks of an individual? => regression analysis
- Identification of segments/factors relevant for pricing? => cluster analysis

**Underwriting efficiency**

- Through which channel should an individual be approached?
  => Classification analysis, cluster analysis

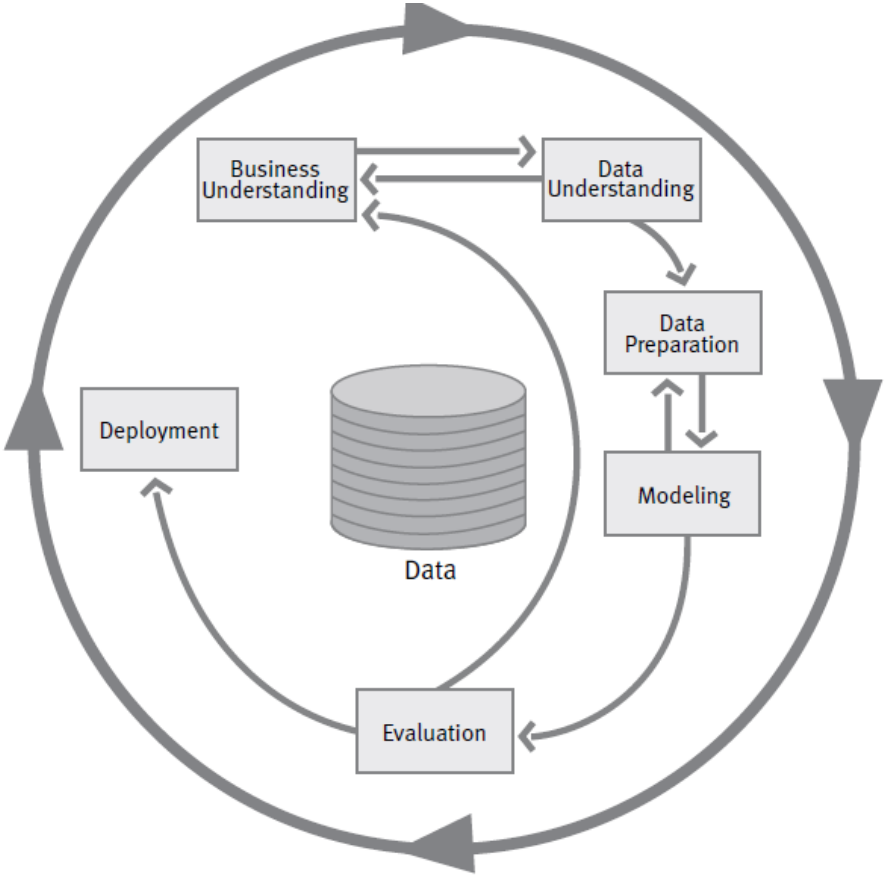
**Customer Life-time Value (CLV)**

- What value will the customer bring to the company?
  => CF projections, regression analysis, Markov chains, …
- More complex problem
- Powerful in combination with other analyses

# Insurance Analytics process

# Insurance Analytics project

- Phases of the CRISP-DM reference model
- The process is not straightforward



# Phases overview

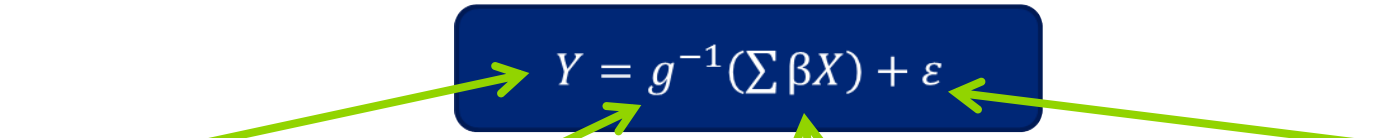
- Components of phases of the CRISP-DM reference model

| Business Understanding | Data Understanding | Data Preparation | Modelling | Evaluation | Deployment |
|---|---|---|---|---|---|
| • Business Objectives<br>• Assess Situation | • Collect Initial Data<br>• Describe Data<br>• Explore Data<br>• Verify Data Quality<br>• 1-way ANOVA<br>• scatter plots, box plots | • Clean Data<br>• Select Relevant Variables<br>• Construct New Variables<br>• Cluster analysis<br>• Identification of outliers<br>• Dimension reduction techniques | • Select Modelling Techniques<br>• Build Model<br>• GLM (logistic regression, …)<br>• CRT<br>• Naïve Bayesian Networks<br>• Neural networks | • Evaluate Results<br>• Asses models<br>• Cross-validation<br>• Cumulative gains,<br>• Lift | • Application of Results<br>• Next steps<br>• Cost-benefit analysis<br>• Optimization |

# Modelling techniques

# GLM - overview

- Based on parametrized statistical model with explicit assumptions




**Observations**
e.g. no. of claims, severity of claims, Lapse/non-lapse,…

**Link function**
Selected a-priory

**Linear predictor**:
Parameters - estimated, Factors

**Random error**
Distribution from exponential family;
Selected a-priori

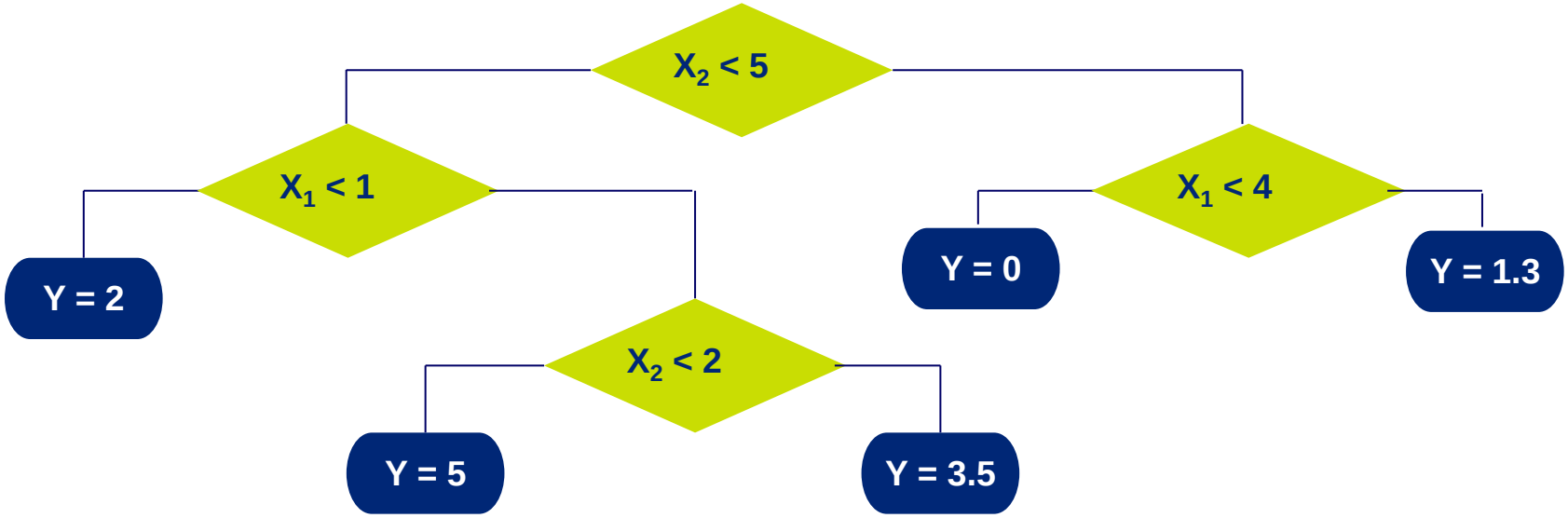
- Parameters estimated using statistical techniques – e.g. maximum likelihood

  => (asymptotic) properties known from general theory

- Includes logistic regression (binary response variable) – useful for classification

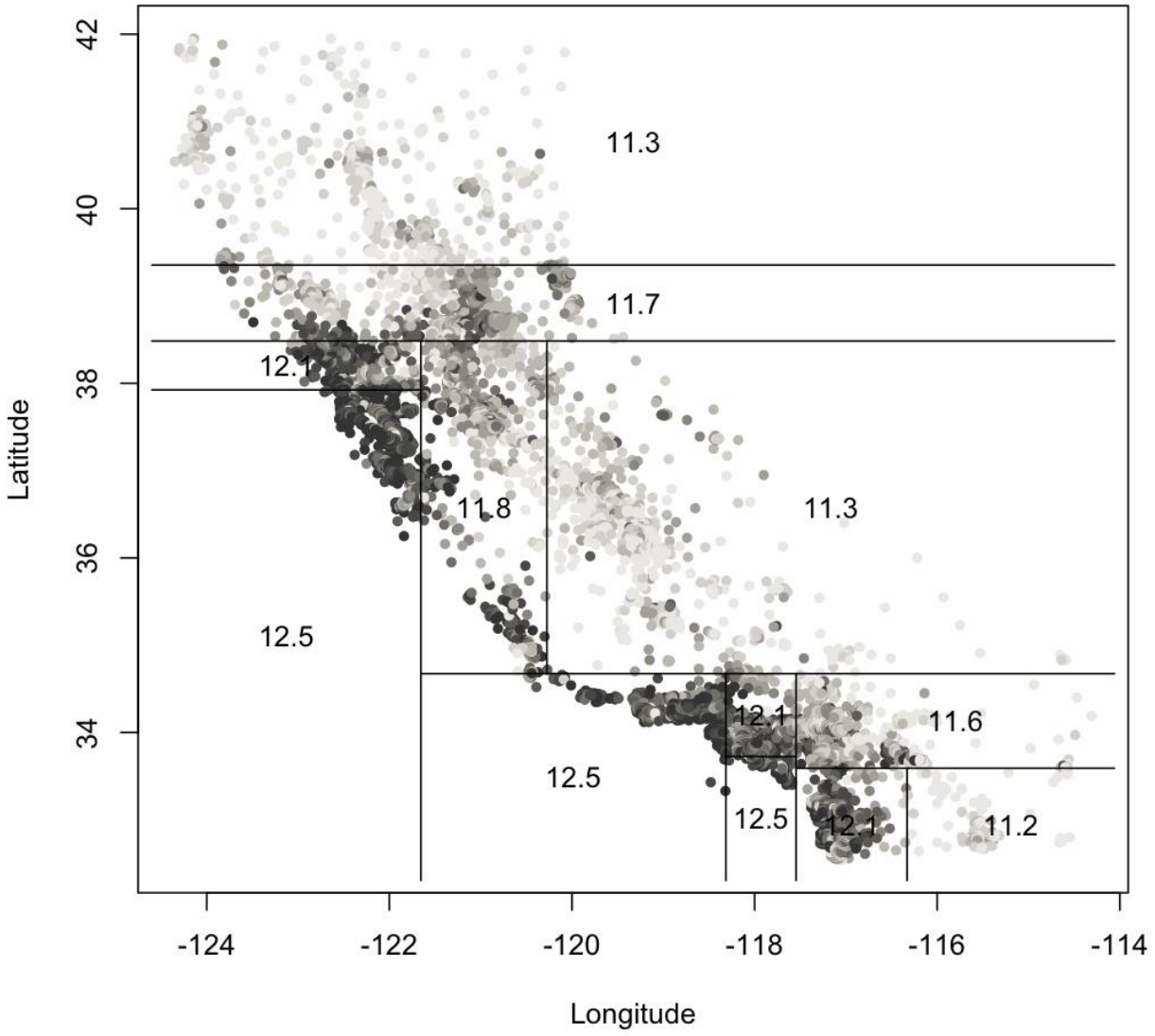
# Classification and regression trees (CRT)

- Tree : $( X_1, X_2, \ldots X_n ) \longrightarrow Y$




- Piece-wise constant function on segments of factor-space
- Classification trees: Y = 0 or Y = 1
  - Predictions in terms of distributions: y = P(Y = 1)

# Regression tree for house prices in California

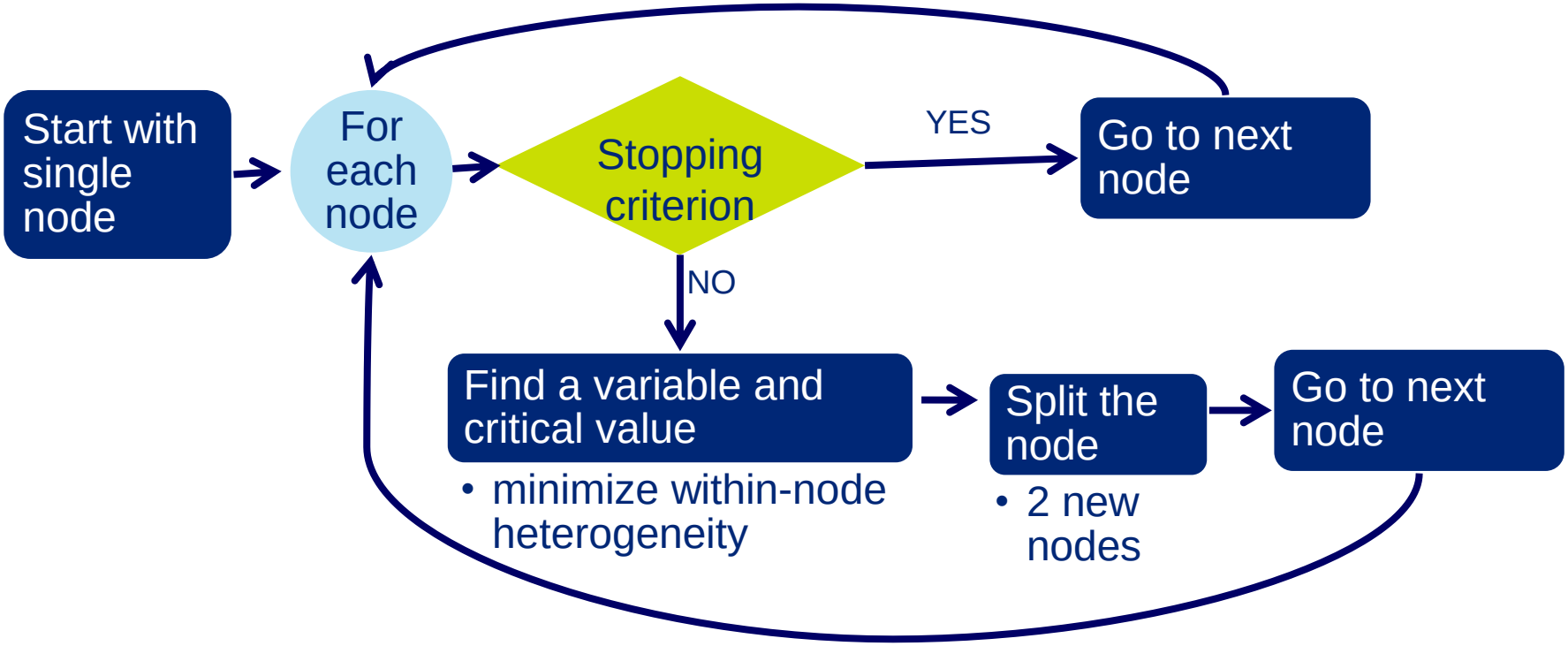



Source: http://www.stat.cmu.edu/~cshalizi/350/lectures/22/lecture-22.pdf

# Growing (learning) CRT

Performed on training data

1. Find good **partitioning**




2. Fit local **models for leaves**: constant function
   - Average over observations in the leave

# Growing CRT

## Stopping criterion

- Number of observations within each child
- Threshold for decrease in heterogeneity

## Measuring heterogeneity

- Various measures => various algorithms

1. Regression trees: **Sum of squared errors**

$$S = \sum_{c \in leaves(T)} \sum_{i \in c} (y_i - m_c)^2, \quad \text{where} \quad m_c = \frac{1}{|c|} \sum_{i \in c} y_i$$

2. Classification trees: **Gini impurity** (⇔ sum of squared errors),

$$I_G = avg_{c \in leaves(T)} \{p_c(1 - p_c)\}, \quad \text{where} \quad p_c = \frac{1}{|c|} \sum_{i \in c} y_i$$

**Entropy**

$$I_E = avg_{c \in leaves(T)} \{-p_c log_2(p_c) - (1 - p_c) log_2(1 - p_c)\},$$

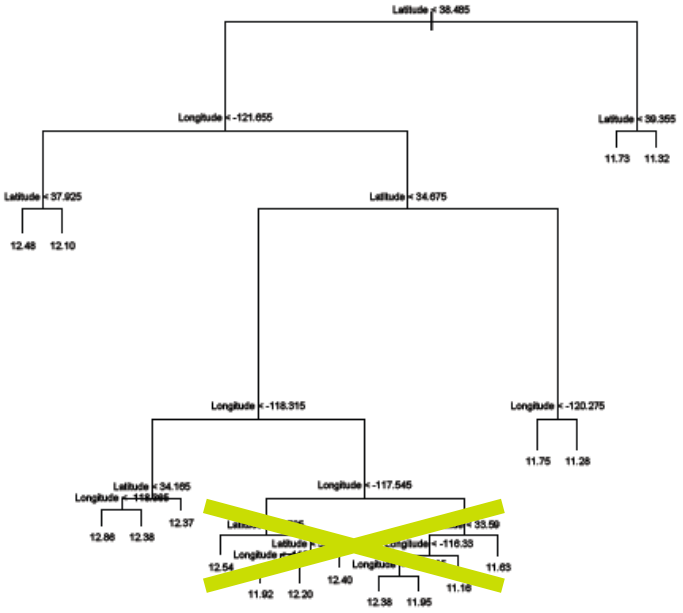
# Pruning CRT

**Problem:** How to set stopping criterion

- **Too strict** => CRT cannot capture local dependencies
- **Too mild** => over-fitting

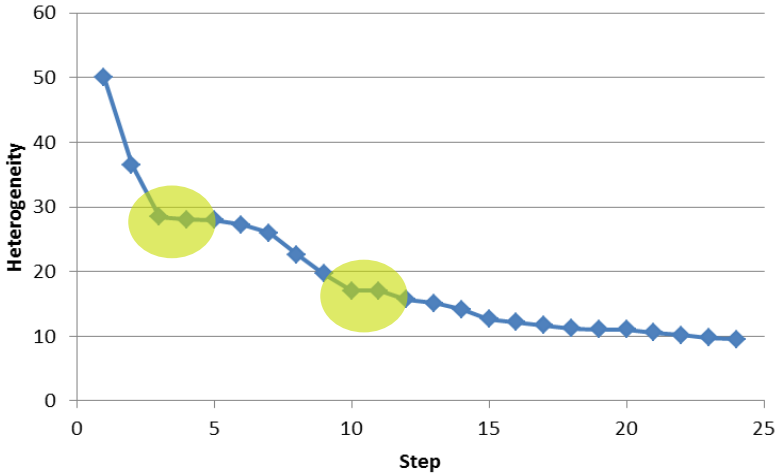
**Way out:** Pruning

1. **Grow large tree:** with mild stopping criterion
2. **Prune the tree:** re-join leaves, which do not decrease heterogeneity




Combine with **cross-validation:**

- 2 sets of records: **training** and **testing**
- Use **training** data to **grow** the tree
- Use **testing** data to **prune** the tree



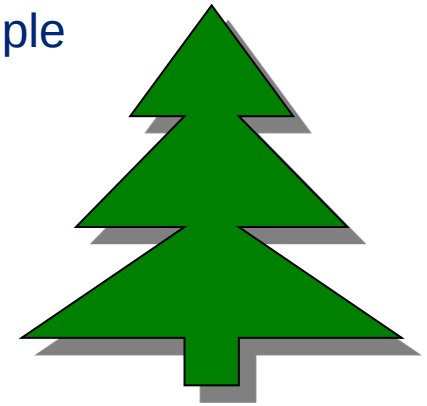
# Uncertainty in CRT

How precise/reliable is prediction from CRT?

1. Error in conditional mean estimate
   - Prediction error assuming the tree is correct
   - Estimate of standard error of mean (under i.i.d. errors)

$$SE_c = \frac{s_c}{\sqrt{|c|}} \qquad \widehat{SE}_c = \frac{\hat{s}_c}{\sqrt{|c|}} = \frac{1}{\sqrt{|c|}} \sqrt{\frac{1}{(|c|-1)} \sum_{i \in c} (y_i - m_c)^2}$$

2. Error in tree fitting
   - How different would the tree be had we drawn different sample
   - **Non-parametric bootstrapping**
     - Draw samples from data with replacement
     - Grow trees on samples => empirical distribution



# Why to use CRT?

## Pros

- Captures local dependencies and interactions
- Making predictions is fast
- Identifies variables important for predictions => **dimension reduction**
- Easy to understand and interpret (white box)
- Able to handle numerical and categorical data
- Robust (not sensitive to assumptions)
- Performs well with large datasets
- Predictions if some variables are missing
- Enables distributional predictions

## Cons

- Not smooth = No sensitivities
- Does not distinguish points
- Difficulties with global patterns
- Learning NP-complete => locally optimal trees

# Naive Bayes Classifiers

Based on:

- Bayes' theorem
- Conditional independence of factors (given the class variable)
  - it is too strong and naive assumption

Probabilistic model

$$P(Y = c|X_1 = x_1, \dots, X_n = x_n)$$

$$= \frac{P(Y = c)\, P(X_1 = x_1, \dots, X_n = x_n|Y = c)}{P(X_1 = x_1, \dots, X_n = x_n)}$$

$$= \frac{1}{Z} P(Y = c) P(X_1 = x_1|Y = c) \dots P(X_n = x_n|Y = c)$$

Parameter estimation

- maximum likelihood = relative frequencies

# Case study: Cross-selling of Caravan insurance

# Business and data understanding

# Business goals

Determination of target customers of X-selling campaign for Caravan insurance

# Data Understanding

## Source

- Public dataset from the Coil 2000 data mining competition
- Downloaded from http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

## Characteristics

- A table with customer data of an insurance company
- 9822 records
- Target variable: Caravan Insurance holder (binary)
- 86 attributes:
  - 43 socio-demographic variables derived from ZIP code
  - 43 variables about premium paid and number of insurance policies
- Only 5% of customers have Caravan Insurance

# Data Exploration

- Single factor analysis, prediction power w.r.t. dependent variable
- Variables categorized => prediction power tested by 1-way ANOVA:

| Summary of 1-way ANOVA results – selected variables with highest F-statistic | | | |
|---|---|---|---|
| **Variable** | **Description** | **F** | **Sig. (p-value)** |
| PPERSAUT | Contribution car policies | 145.393 | 4.36E-33 |
| APERSAUT | Number of car policies | 123.607 | 1.98E-28 |
| ATOTAL | Total number of insurance ADJ | 104.262 | 2.84E-24 |
| PTOTAL | Total insurance contribution ADJ | 78.093 | 1.28E-18 |
| APLEZIER | Number of boat policies | 65.758 | 6.16E-16 |
| PWAPART | Contribution private third party insurance | 56.454 | 6.62E-14 |
| MKOOPKLA | Purchasing power class | 54.066 | 2.21E-13 |
| MINKGEM | Average income | 47.725 | 5.43E-12 |
| MOPLEDUC | Education level ADJ | 46.977 | 7.92E-12 |
| AWAPART | Number of private third party insurance | 46.877 | 8.34E-12 |
| MSKSOCIAL | Social class ADJ | 37.892 | 7.98E-10 |
| MINCOME | Average income ADJ | 31.339 | 2.27E-08 |
| MRELGE | Married | 29.271 | 6.54E-08 |
| PPLEZIER | Contribution boat policies | 28.946 | 7.73E-08 |

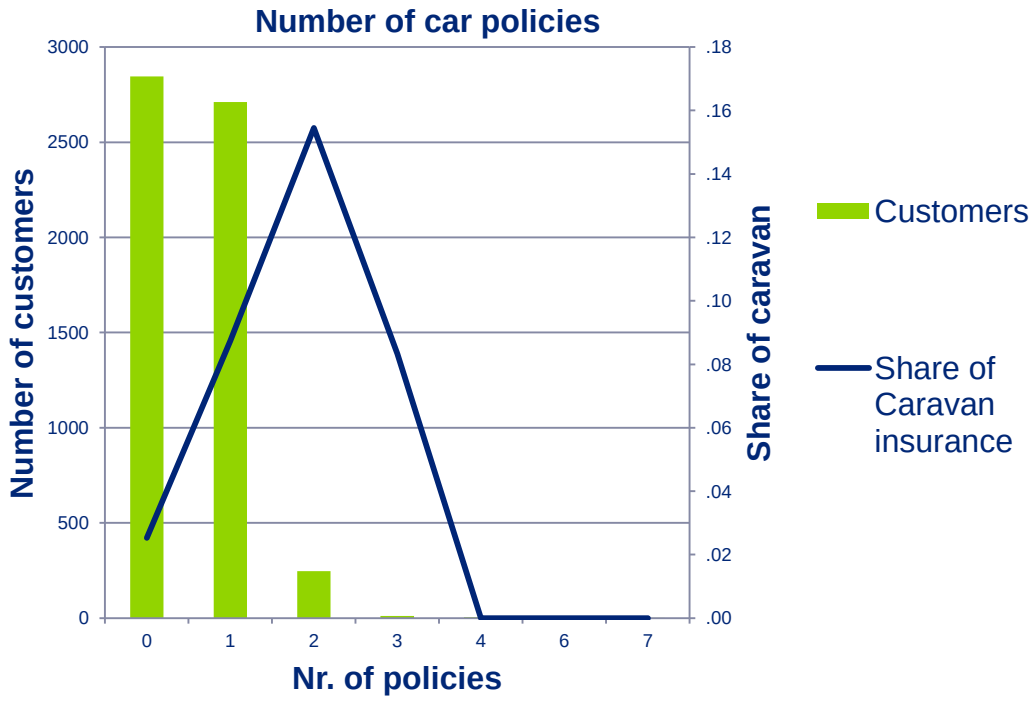
- Factors ordered by F-statistic: the higher the F-statistic, the better prediction potential of the factor
- No information about the “sign” of the dependency, the dependency need not be monotonous
- Does not capture correlations

# Data Exploration

- More detailed analysis of selected individual factors
  - Selected based on **business feeling** and **1-way ANOVA**
  - Main expected factors to determine propensity to buy Caravan insurance are:
    - Ownership of  car(s), wealth (purchasing power), risk aversion (propensity to buy insurance)

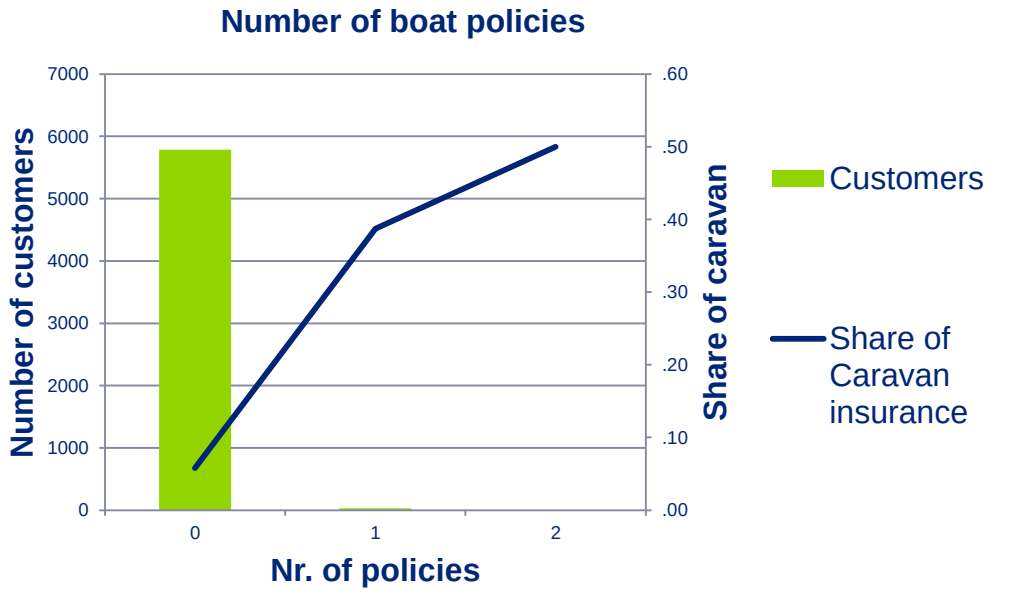
## Example of single factor analysis:

- Car insurance policies:
  - More policies => higher propensity to buy Caravan insurance
  - Classes with low number of customers irrelevant



# Data Exploration

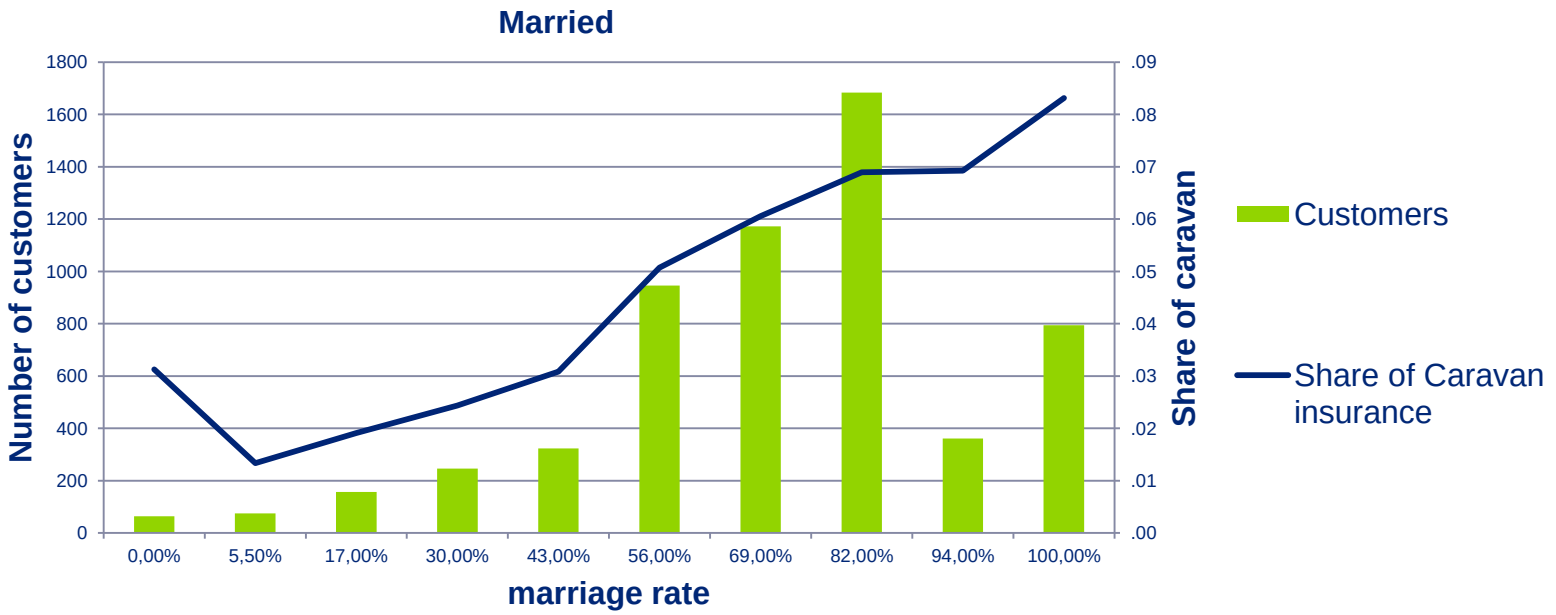
- Boat insurance
  - Having a boat insurance significantly increases propensity to have Caravan insurance
  - Only few customers have boat insurance




- Purchasing power class
  - Prediction power of factor “purchasing power class” appears rather limited (except for class 7)

# Data Exploration

- Family status:
  - Higher chance of being married (derived from ZIP code) increases chance to buy caravan insurance
  - Current marital status (not available in current data) might be useful factor

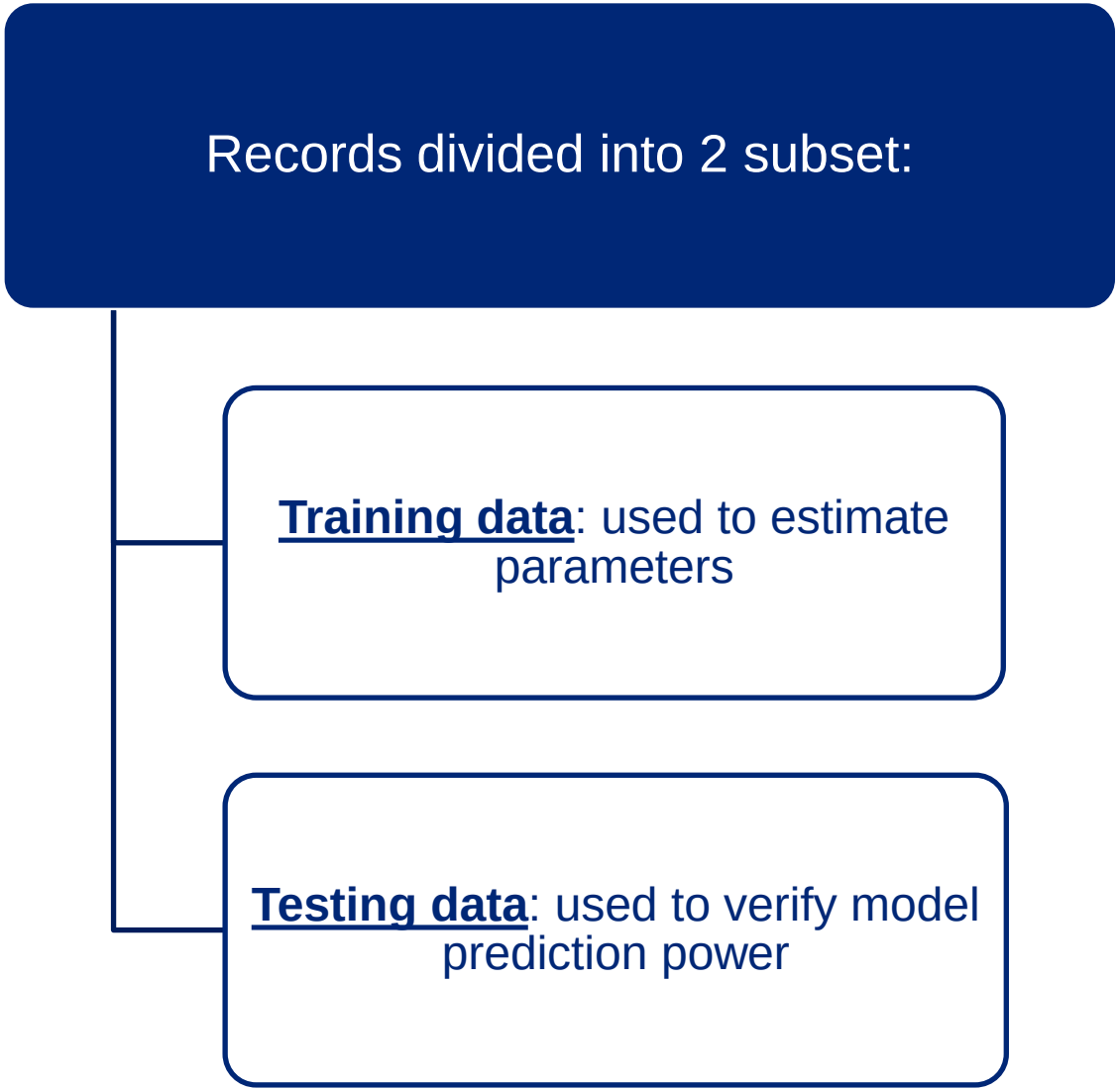


# Data Preparation

- Can by very time-demanding
- Important for successful modelling
- Includes:
  - Data cleansing;
  - Derivation of new factors
  - Use of external sources
  - etc.

# Modelling

# Cross-validation



# Logistic regression

- Special type of GLM
  - Bernoulli 0-1 valued response variable: P(Y=1) = p, P(Y=0 ) =1- p
- Regression equation:

$$log\frac{p}{1-p} = \sum_{i=1}^{K} B_i \times x_i$$ where $B_i$ = parameter; $x_i$ = variable

- Based on forward stepwise selection and single factor analysis, we selected following variables for the model

| Variable | Description |
|---|---|
| PPERSAUT | Contribution car policies |
| MKOOPKLA | Purchasing power class |
| MRELGE | Married |
| MOPLEDUC | Education level ADJ |
| PTOTAL | Total insurance contribution ADJ |
| APLEZIER | Number of boat policies |

# Logistic regression – dummy variables

- Recoding of categorical variable KOOPKLA (purchasing power class) into several dummy (0-1) variables
  - Enables use of nominal/categorical variables
  - Enables modelling non-monotonous dependencies

| Categorical Variables Coding | | | | | | | |
|---|---|---|---|---|---|---|---|
| Factor value | Dummy variable | | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# Logistic regression

- Regression equation:

$$log\frac{p}{1-p} = \sum_{i=1}^{K} B_i \times x_i$$

where $B_i$ = parameter; $x_i$ = variable

- Parameter interpretation:
  - B … estimated parameter value
  - Sig…significance level (p-value) of the hypothesis “parameter = 0”. Low values indicate significant variables (i.e. sig = probability that real parameter is zero)
  - Exp(B)…percentage change of odds (p/1-p) by changing the corresponding variable by 1 unit
    - Example (see below): Increase in number of boat policies by 1 increases odds (ratio of those with caravan insurance to those without caravan insurance) 7.3 times – compare to single factor analysis.

| Parameter estimates | | | | |
|---|---|---|---|---|
| **Variable** | **Description** | **B** | **Sig.** | **Exp(B)** |
| APLEZIER | Number of boat policies | 1.99262 | 0.00000 | 7.33469 |

# Logistic regression – parameter estimates

| MLE Parameter estimates | | | | |
|---|---|---|---|---|
| **Variable** | **Description** | **B** | **Sig.** | **Exp(B)** |
| MKOOPKLA | Purchasing power class | | 0.00106 | |
| MKOOPKLA(1) | | -0.44840 | 0.16598 | 0.63865 |
| MKOOPKLA(2) | | -0.30545 | 0.38822 | 0.73679 |
| MKOOPKLA(3) | | -0.24723 | 0.29789 | 0.78096 |
| MKOOPKLA(4) | | -0.13853 | 0.58047 | 0.87064 |
| MKOOPKLA(5) | | -0.09847 | 0.72481 | 0.90622 |
| MKOOPKLA(6) | | 0.06729 | 0.76930 | 1.06961 |
| MKOOPKLA(7) | | 0.60826 | 0.00710 | 1.83723 |
| PPERSAUT | Contribution car policies | 0.00046 | 0.00000 | 1.00046 |
| MRELGE | Married | 0.01052 | 0.00054 | 1.01057 |
| MOPLEDUC | Education level ADJ | 0.36838 | 0.00036 | 1.44539 |
| PTOTAL | Total insurance contribution ADJ | -0.00008 | 0.06396 | 0.99992 |
| APLEZIER | Number of boat policies | 1.99262 | 0.00000 | 7.33469 |
| Constant | | -4.72476 | 0.00000 | 0.00887 |

- Being in purchasing power class 7 significantly increases probability of caravan ins. compared to base class 8, other classes rather decrease the probability, but the estimates are not significant (compare to single factor analysis)
- Increase in contributions to car insurance increases probability of caravan ins. The sensitivity to change in contributions by 1 monetary unit is not high. However, the parameter is significant and should stay in the model.
- Higher probability of being married increases probability of caravan insurance.
- Higher education level increases probability of caravan insurance.
- Higher insurance contributions slightly decrease probability of caravan ins, which is contraintuitive (compare to single factor analysis. However, the parameter is insignificant. => Should be excluded from the model?
- Having boat policy => high probability of having caravan insurance.

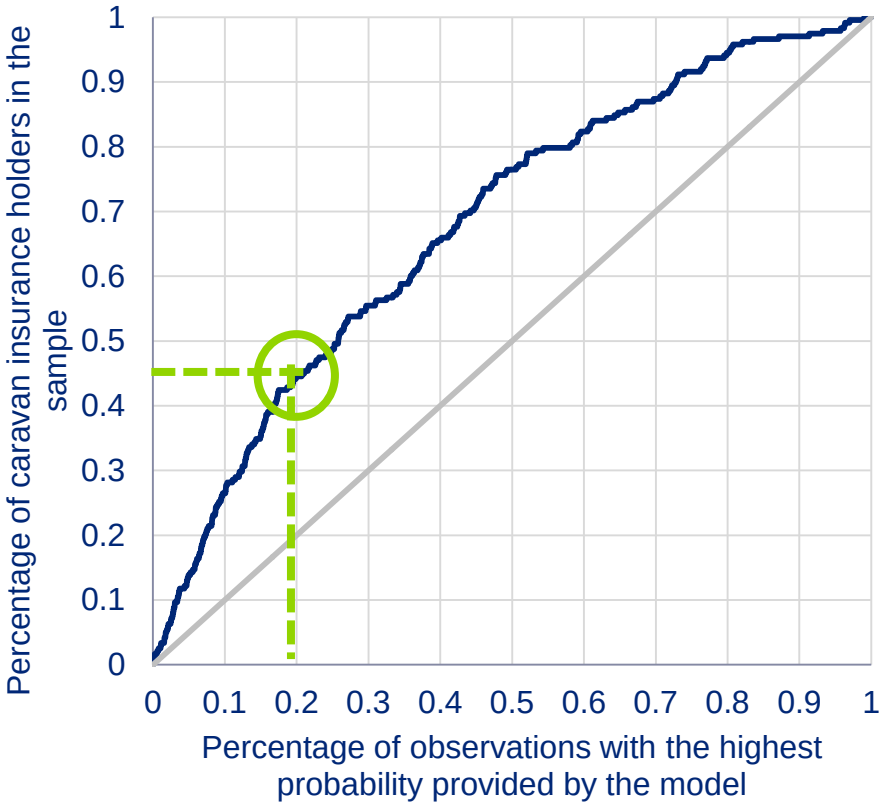
# Logistic regression – diagnostics of classification

- Predictions on **testing set**
- Correct predictions = 3629 + 26 = 3655
  - If we predicted 0 => 3762 correct predictions (but useless)
- Prediction = 1 => 26 correct, 133 misclassified
- Seemingly poor classifier – due to unbalanced data (only 6% caravan ins. holders)

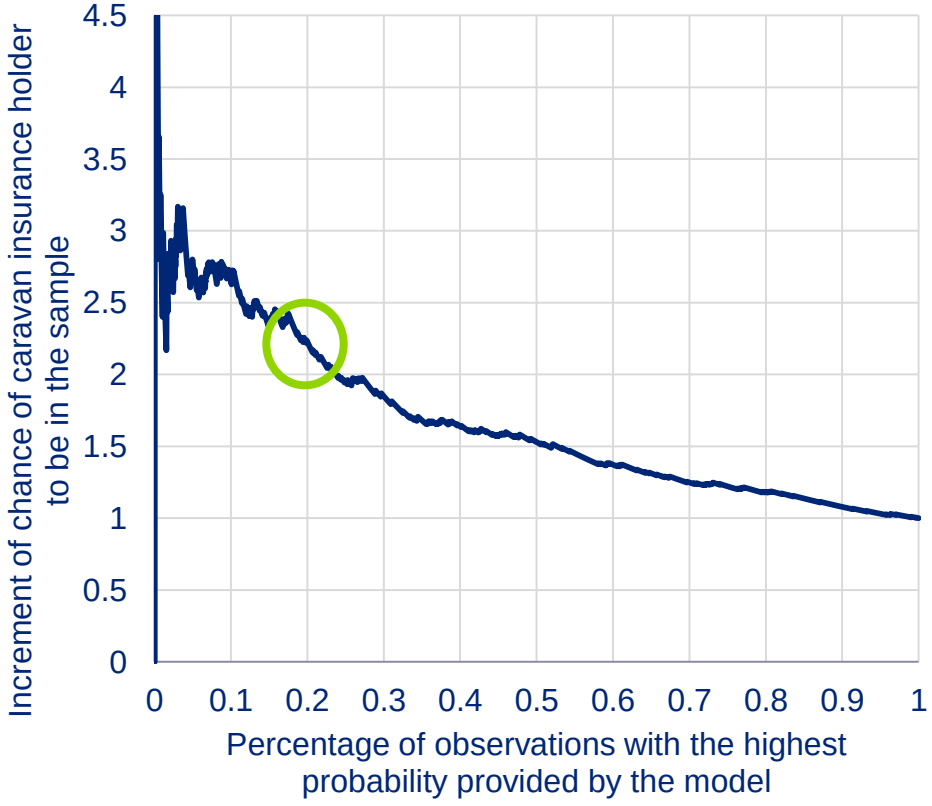
| | | Prediction | | Sum |
|---|---|---|---|---|
| | | 0 | 1 | |
| Real value | 0 | 3629 | 133 | 3762 |
| | 1 | 212 | 26 | 238 |
| Caravans share | | 6% | 16% | 6% |

# Logistic regression – diagnostics of probabilities

## Cumulative Gains



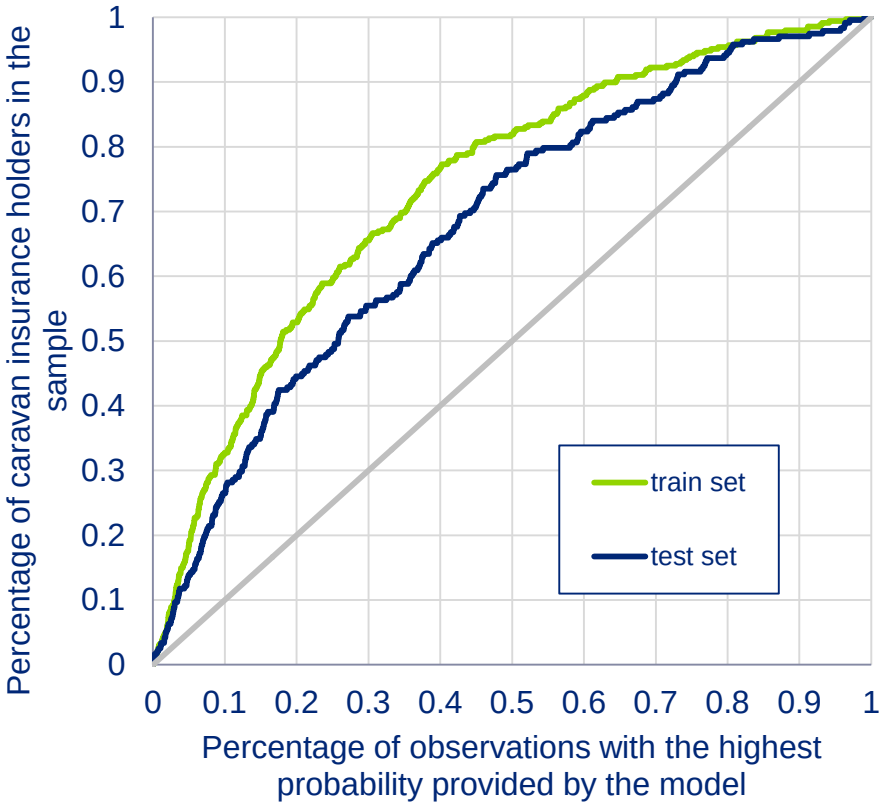

## Lift



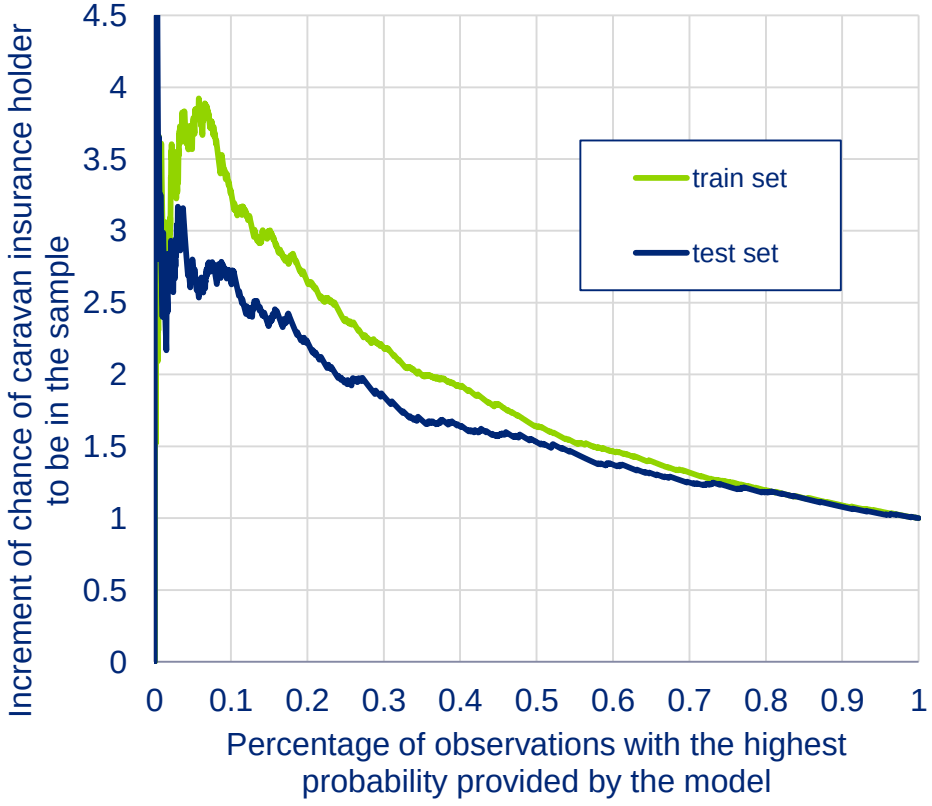

- Cumulative Gains: In 20% of all observation (with the highest probability), there are 45% of all caravan insurance holders.
- Corresponding Lift: for best 20% of observation the method is 2.5x better then random selection

# Logistic regression – diagnostics of probabilities 2

## Cumulative Gains



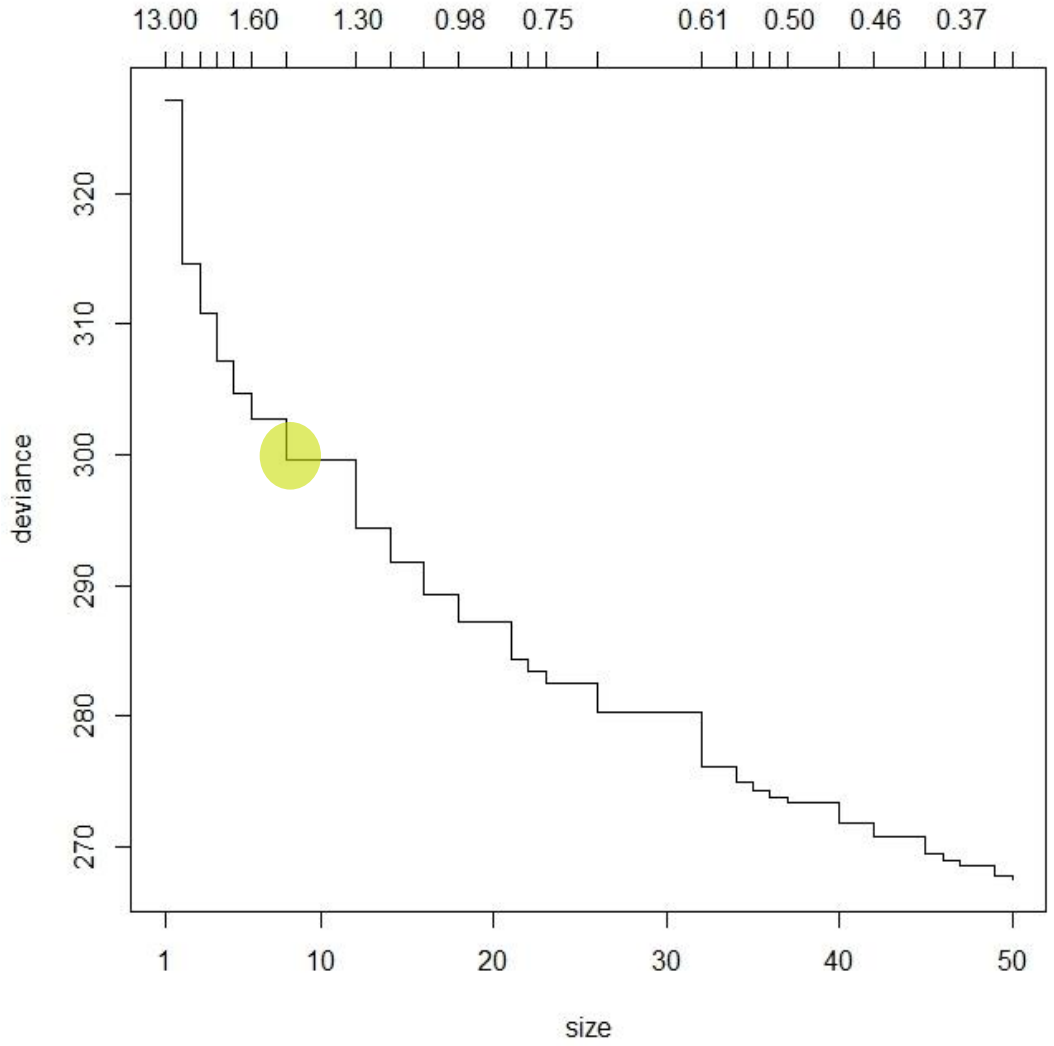

## Lift




- Cumulative Gain/Lift slightly better for training set => slight **over-fitting**

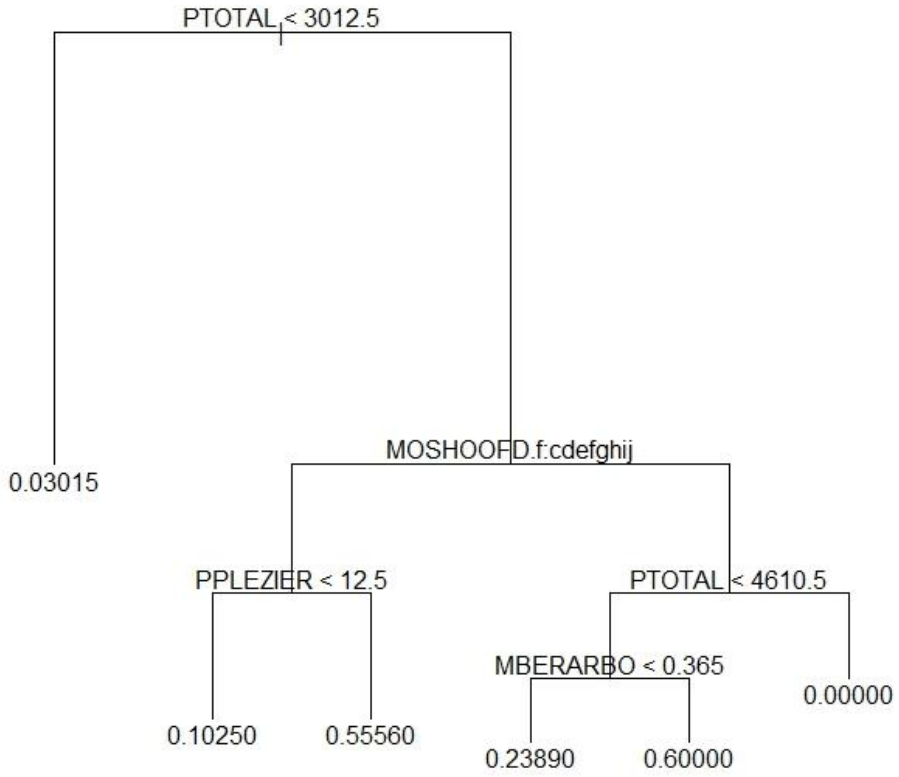
# Classification tree – growing

- Pre-selection of variables not needed
- Graph of the decrease in deviance (Gini impurity) by adding nodes



# Classification tree - pruned

- Based on graph, originally selected 20 nodes => over-fitting
- Nr. of nodes reduced to 6 => worse Lift for training set, but better Lift for testing set



# Classification tree – selected factors

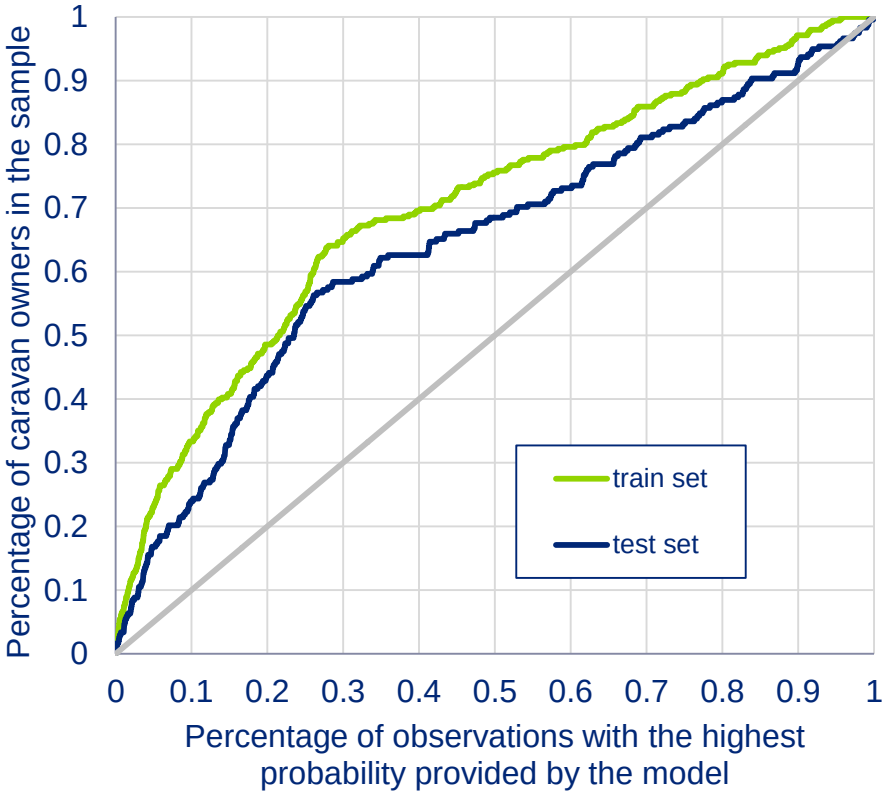
| Variable | Description |
|---|---|
| PTOTAL | Total insurance contribution ADJ |
| MOSHOOFD | Customer main type (L2) |
| PPLEZIER | Contribution boat policies |
| MBERARBO | Unskilled labourers |

| Customer main type | |
|---|---|
| L2 | Label |
| a | Successful hedonists |
| b | Driven Growers |
| c | Average Family |
| d | Career Loners |
| e | Living well |
| f | Cruising Seniors |
| g | Retired and Religeous |
| h | Family with grown ups |
| i | Conservative families |
| j | Farmers |

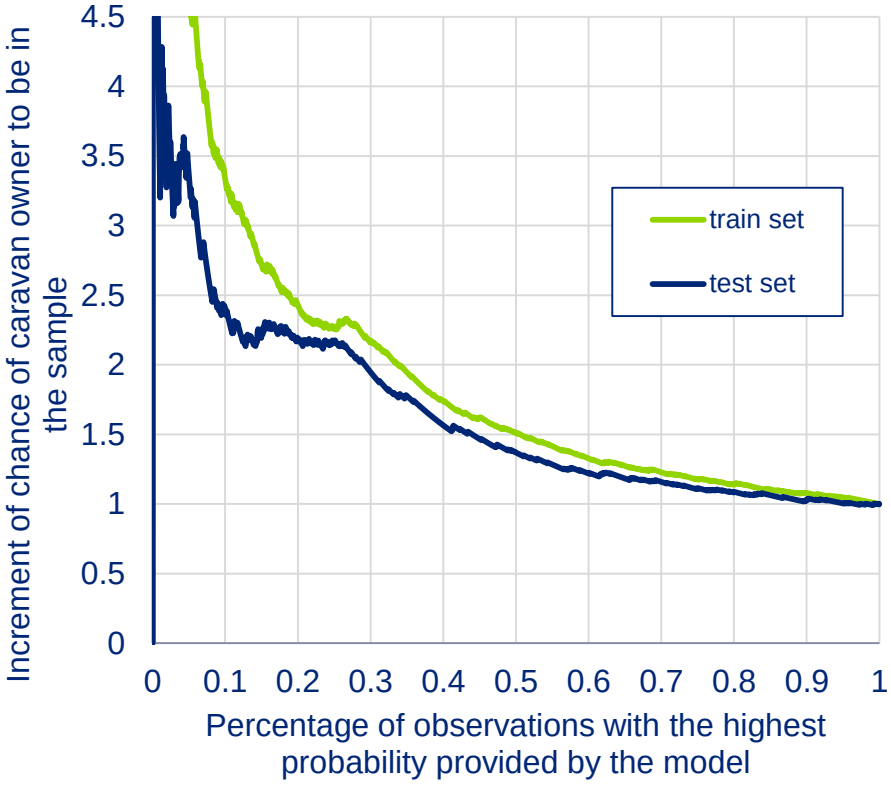
- Selected variables to a certain extend correspond to single factor analysis.
- Number of observations in terminal nodes seem to be unbalanced

# Classification tree - diagnostics

## Cumulative Gains



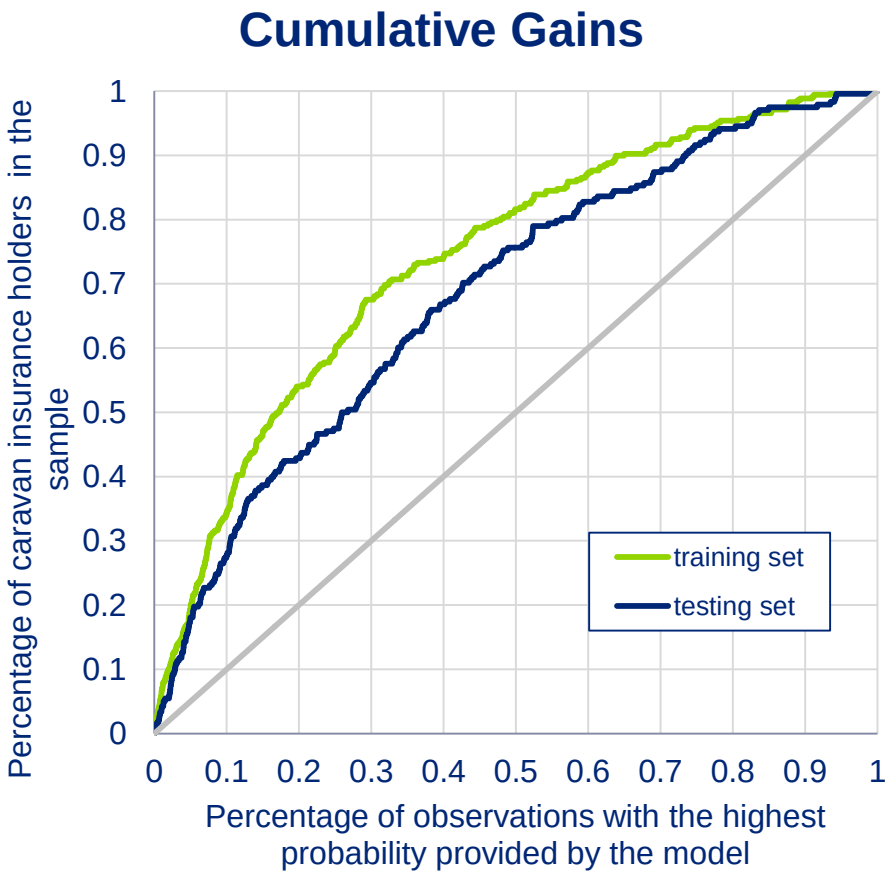

## Lift

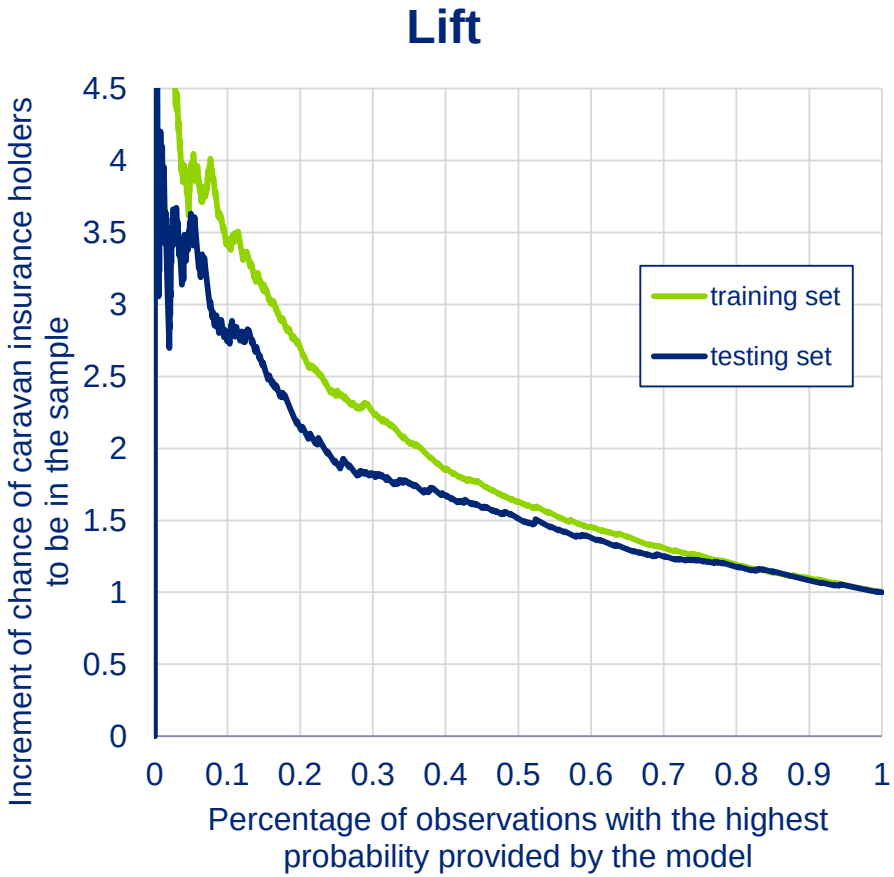



- The tree usable for up to 30% of the observations
- The identification of remaining caravans seems comparable to random selection (due to unbalanced leaves)

# Naive Bayes Classifier

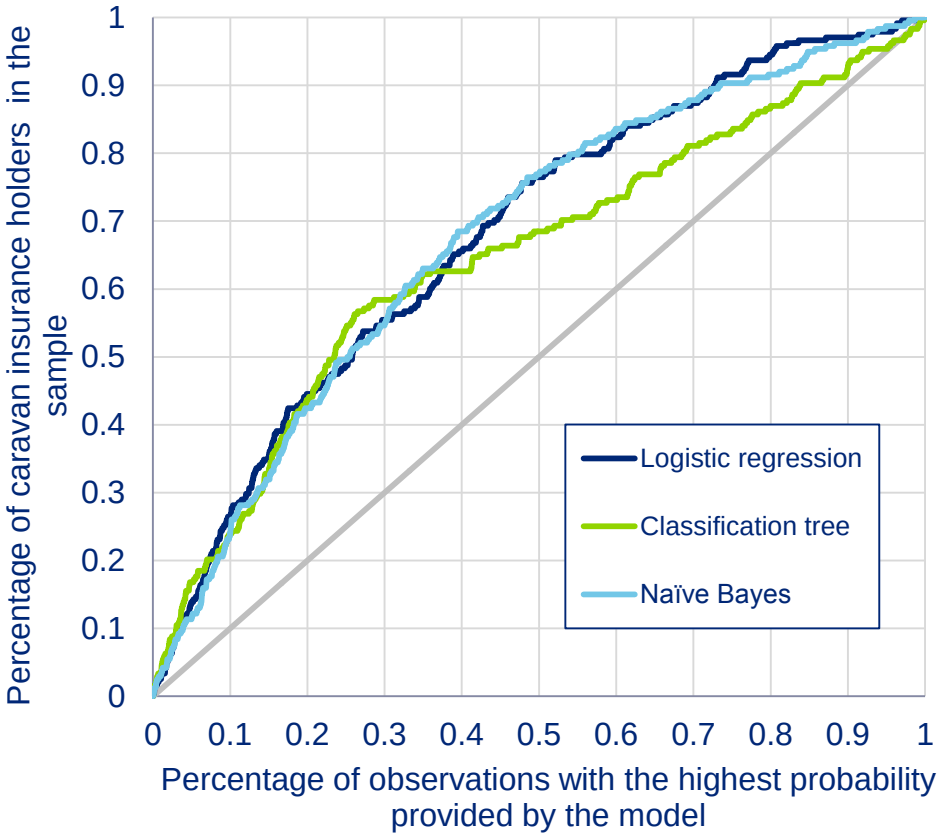
- Selected same factors as in logistic regression



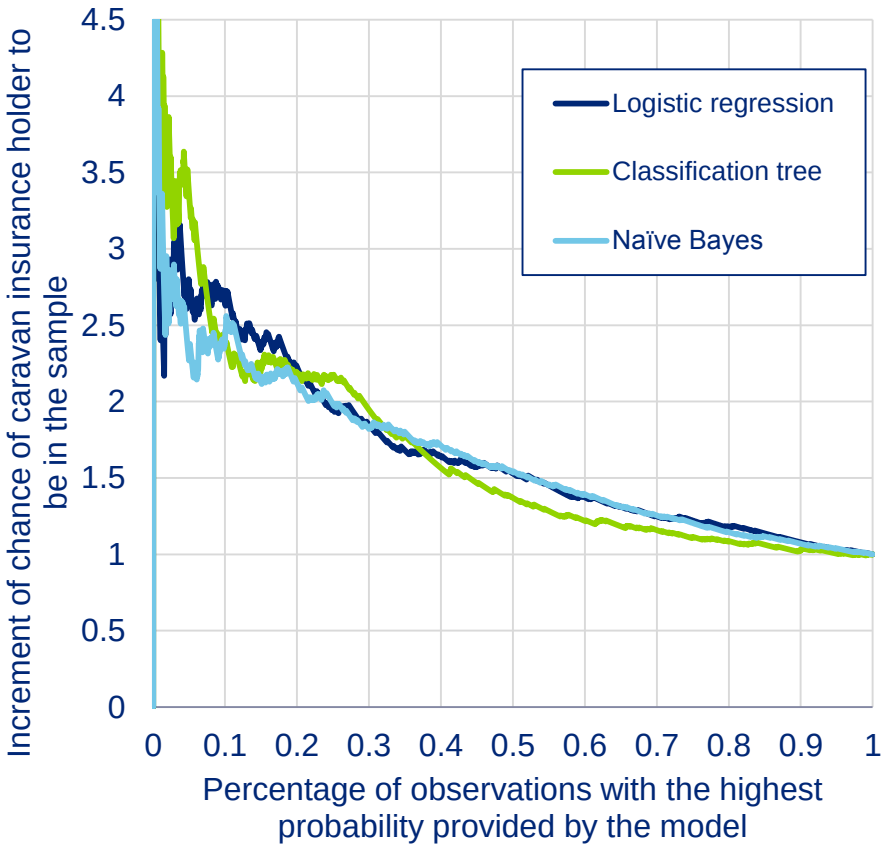



# Comparison of the methods on testing set

## Cumulative Gains




## Lift




- Up to 30% - 40% of observations all the methods give similar results
- For the remaining part logistic regressions and naïve Bayes are better than classification tree

# Cost-benefit analysis

# “Ex-post” cost analysis

## What are the costs to win the desired number of caravan insurance?

- Based on logistic regression
- Dependency of costs on the number of caravan insurance buyers **within testing set**
- Comparison of costs without analysis (blue line) and with analysis (green line)
- Assumptions
  - Initial costs to start the campaign = 5 000 €
  - Costs per customer = 4 €
  - Success rate in the population = 6% (computed from the testing data set)

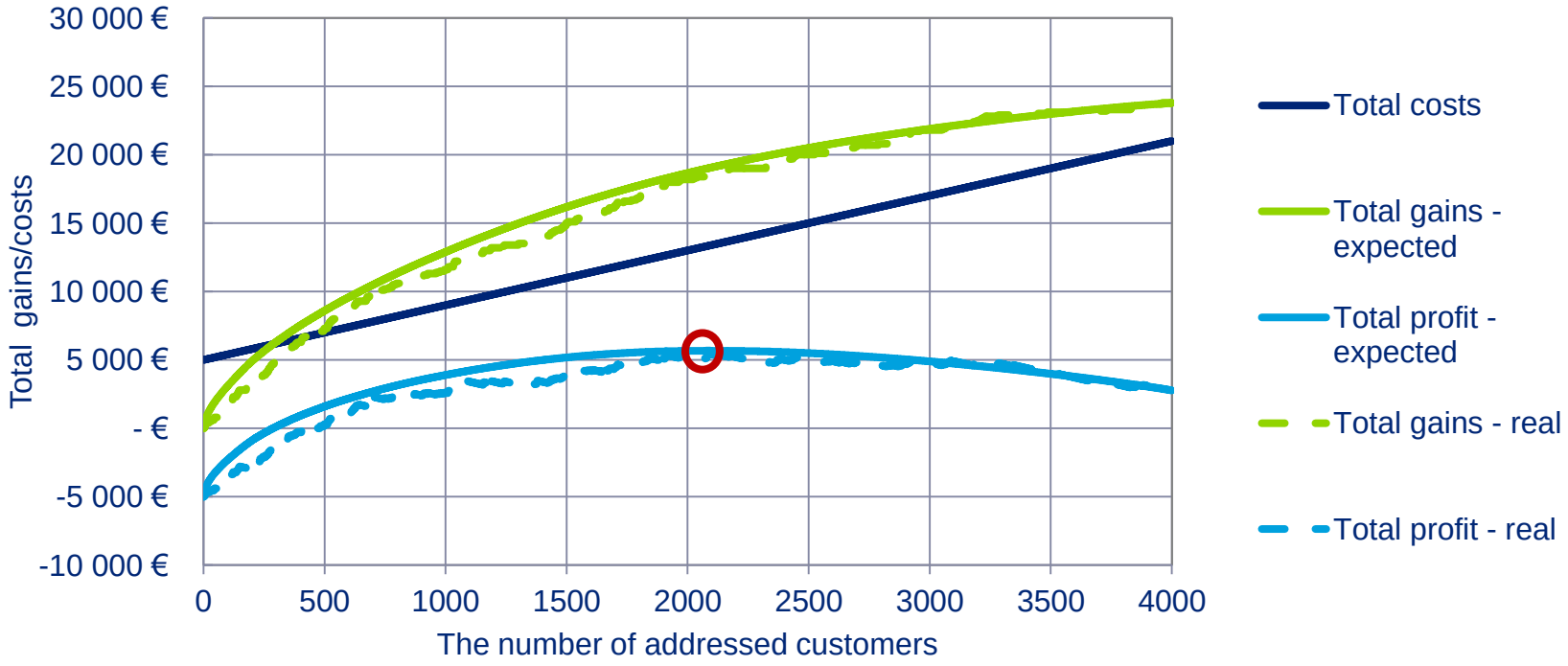

Example: Suppose, we want to sell 100 caravan insurance

- Without analysis, the costs would be 11 704 €
- With analysis, the costs would be only 7 780 €

# Profit analysis

## What is the profit when addressing desired number of customers based on the analysis?

- Dependency of profit on the number of addressed customers **within testing set**
- Assumptions
  - Initial costs to start the campaign = 5 000 €
  - Costs per customer = 4 €
  - Profit from 1 customer = 100 €
- Maximum profit is 5 460 €
- The maximum profit is reached at 2085 addressed customers



# Thank you for your attention.

# Questions?