TECHNISCHE UNIVERSITÄT
BERGAKADEMIE FREIBERG
Die Ressourcenuniversität. Seit 1765.

IAMG 2017, Fremantle

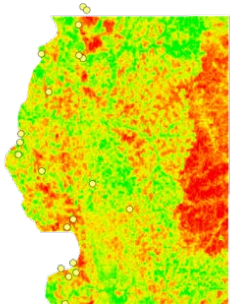
Logistic Regression Model Selection Strategy for Prospectivity Modeling with Large Datasets

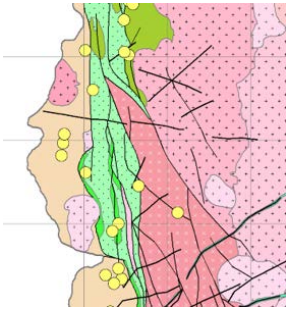
Michael Eiermann*, Samuel Kost*, Oliver Rheinbach*, Helmut Schaeben*

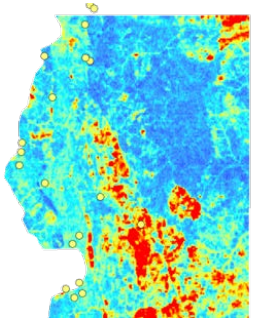
*Technische Universität Bergakademie Freiberg, kosts@mailserver.tu-freiberg.de

Outline

- Introduction
- Basics
- Model Selection Strategy
- Experiments
- Conclusion







Introduction - Prospectivity Modeling

Goal

- Large conditional probability of target event

How?

- Collect data → design “good“ Model

Methods

- Artificial Neural Nets
- Random Forrest
- Support Vector Machines
- **Logistic Regression**

Introduction – Existing Problems

Big Data

3d models

- Lots of information
- Lots of different features

Black Box Methods

- Good prediction accuracy
- Small information gain
- No understanding of problem
- No or small information about used features

Basics – Logistic Regression

- Linear regression model

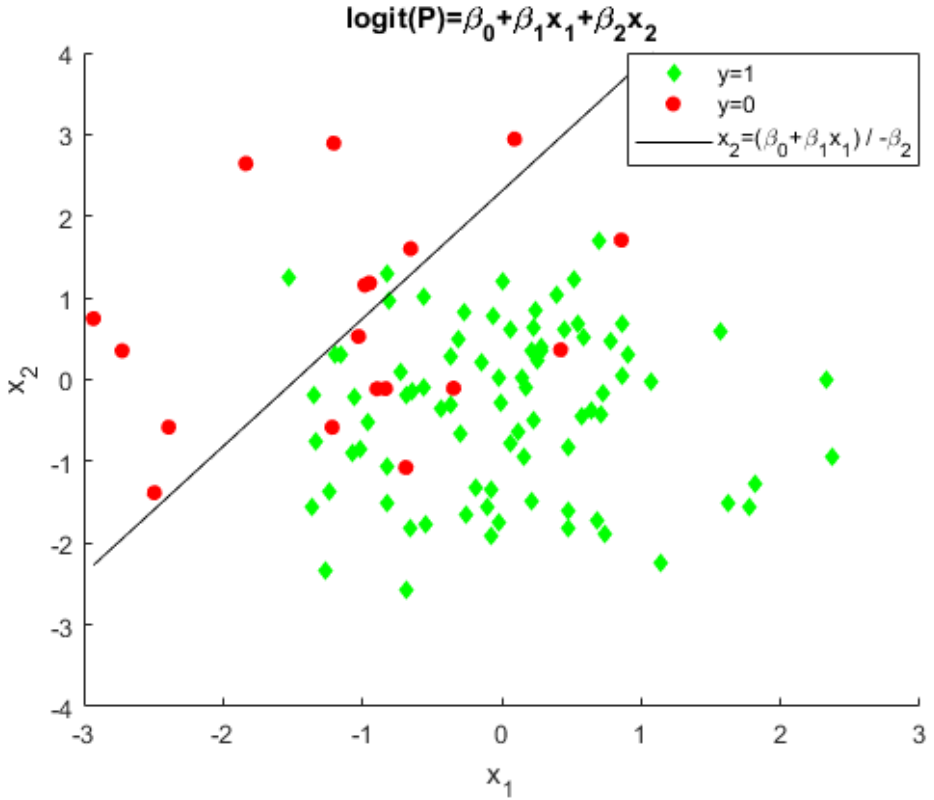
$$\hat{y} = \beta_0 + \sum_{j=1}^{m} x_{ij}\beta_j$$

- Apply logit transformation

$$ln\frac{P(y_i=1|x_i)}{1-P(y_i=1|x_i)} = \beta^T x_i$$

- Apply logistic function

$$P(y_i = 1|x_i) = \frac{\exp(\beta^T x_i)}{1+\exp(\beta^T x_i)}$$

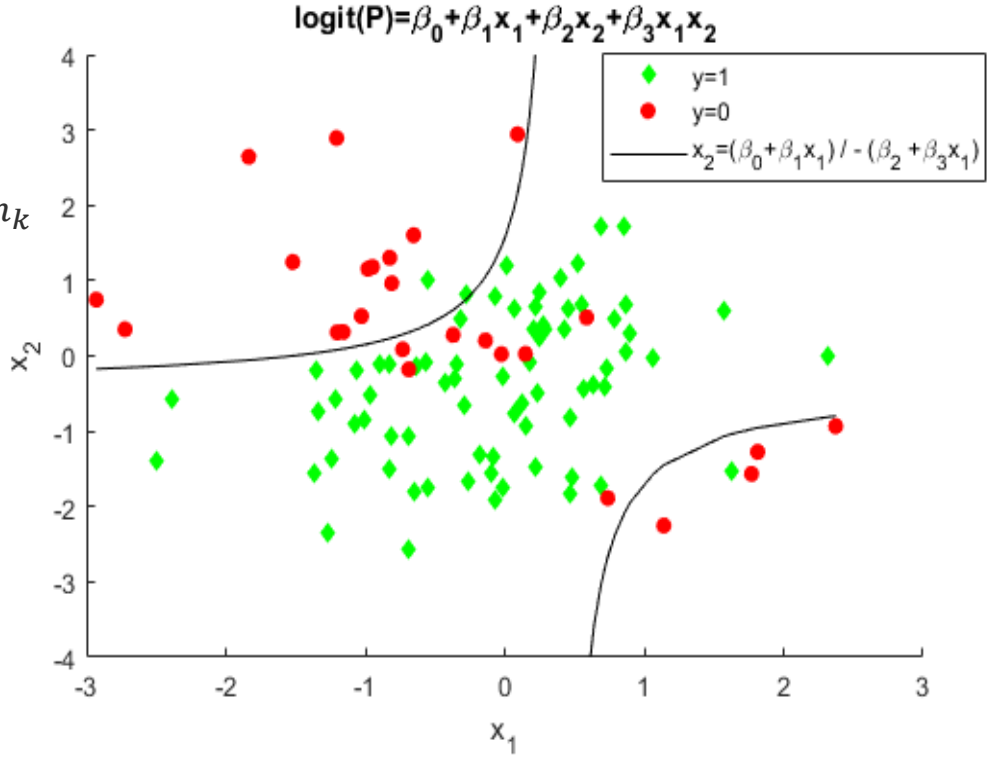


Basics – Logistic Regression Interactions

- Additional dependencies between features

$$\text{logit}(P) = \beta^T x_i + \sum \beta_{m_l \ldots m_k} x_{m_l} \otimes \ldots \otimes x_{m_k}$$

- All kind of non-linearities possible (two-fold,three-fold,squares,cubes)
- Results in non-linear classification border
- Classification of more complex data



Basics – Endogenous Sampling

- Rare events (< 1% events)
- Positive targets more information than negative targets
- Select all events + random sample of non-events to have balanced subsample

➡ Reduction of computational costs!

Corrections[1]

- Solution of weighted ML
- Rare events underestimated

→ Calculate bias correction

1, Gary *King* and Langche *Zeng*. 2001. “Logistic Regression in Rare Events Data.”

Basics– Bayes Information Criteria (BIC)

- Measures quality of models

$$BIC_M = \ln(n)\, k - 2\ln(\hat{L}_M)$$

- Trade-off between goodness of fit and complexity

- The smaller the BIC the better

- Punishment of complex models greater than with Akaike‘s Information Criteria

$$AIC_M = 2\mathrm{k} - 2\ln\left(\hat{L}_M\right)$$

Model Selection Strategy

- Data is normalized
- Divide data set into three parts (Training, Validation, Testing)

- **Two parts**
  1. Rough feature selection

     Using statistic significance (p-value)

  2. Fine feature selection

     Using Bayes Information Criteria

Model Selection Strategy

First part:

1. Include all non-linearities (i.e. Interactions)
2. Endogenous sampling
3. Run LR with correction
4. Discard features with $p > \alpha$
5. Repeat from 2. until all features are significant ($p < \alpha$)

Note:

- The p-Value will approach zero as sample size increases
- Every feature will become significant with sufficiently large sample

But:

- Can be used for rough selection

Model Selection Strategy

Second part:

1. Endogenous Sampling
2. Run LR with correction
3. Calculate BIC of current model on validation set
4. Loop through all remaining features
    - Compute BIC_j of Model **without** feature j on validation set
5. Discard feature j for which $BIC - BIC_j$ is highest and positive
6. Repeat from 1. until no improvement in BIC can be achieved

Experiments – Synthetic Data

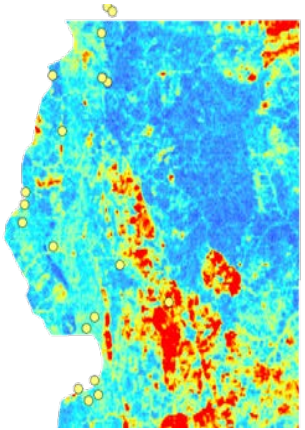
- 10^7 data points with 20 features: $x_i \sim N(0,1)$
- 5 random interactions: $x_i x_j = x_i \cdot x_j$
- True parameter choosen randomly: $\beta \in [-3,3]$
- Intercept $\beta_0 = -20$ to create rare event
- y binomial distributed
- Gives approximately 20.000 positive events (0.2%)

Experiments – Synthetic Data

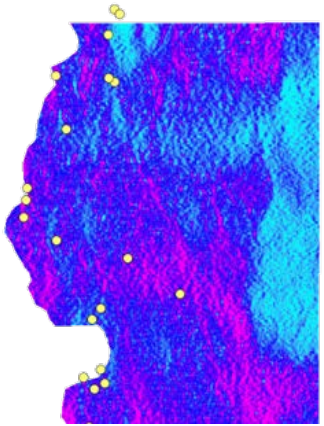
- Data matrix $X \in \mathbb{R}^{10^7 \times 21}$ (intercept included)
- Divide into training, validation, testing (70,15,15)
- Add all squares, two-fold and three-fold interactions
- Giving a total of **1371 features**
- Run algorithm:
    - **First part:** 4-9 loops: leaving 26 - 100 features
    - **Second part:** leaves only 20-26 features (all correct)
        - Discards some correct features with low β without loss of prediction quality

Experiments – Real Data: Search For Gold

- Data provided by Beak Consultants GmbH, Freiberg
- Almost 1 million data points – 4000 positives (0.4%)
- 17 features (geological ,geochemical, geophysical)



Potassium



Uranium

Experiments – Real Data: Search For Gold

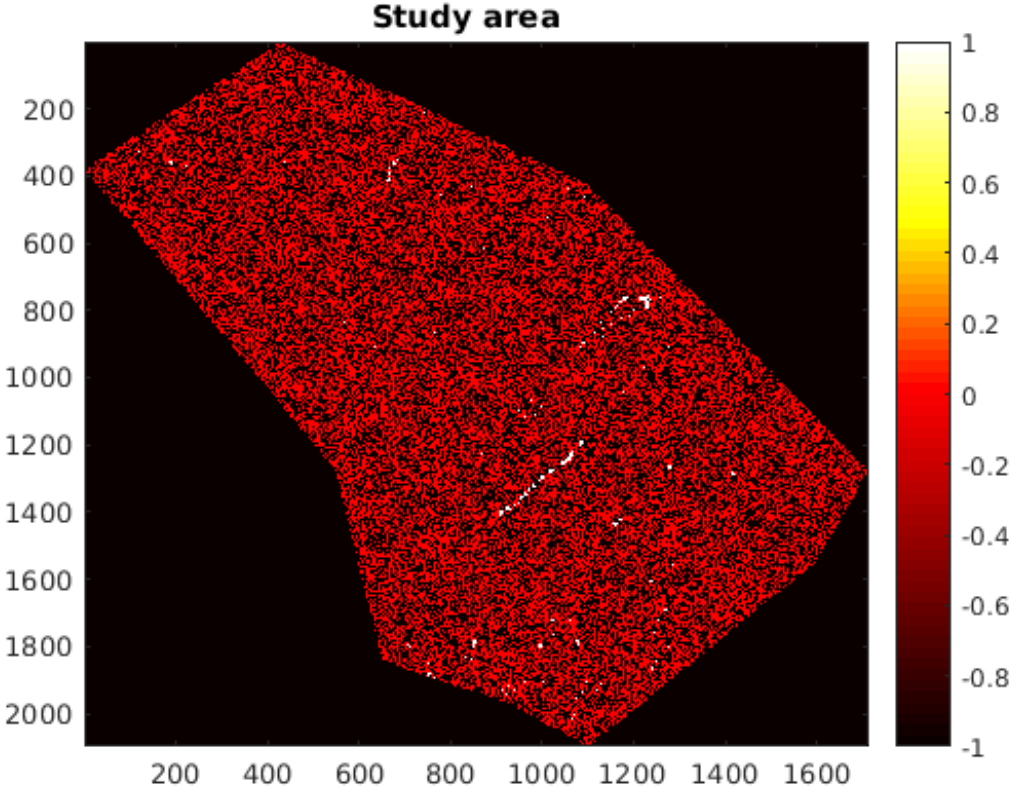
From 17 single features the procedure selected

- 5 single features
- 1 square of a feature
- 3 two-fold interactions
- 1 three-fold interaction

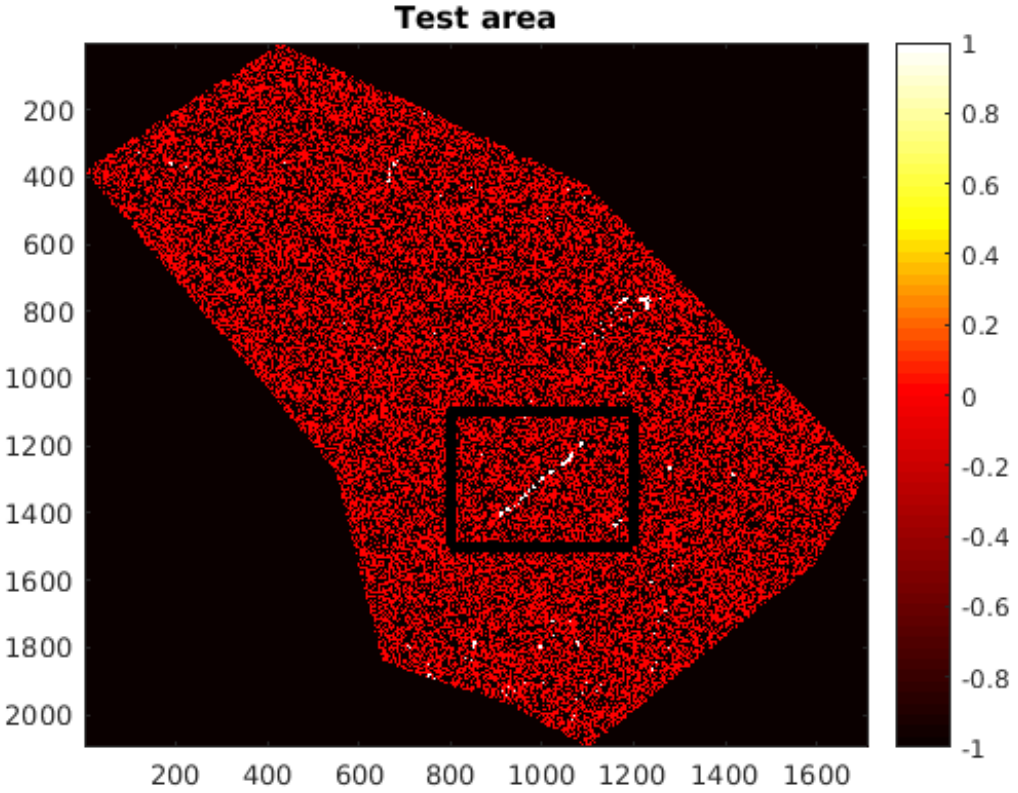
giving a total of **10 variables**

- Prediction results are way better than with simple logistic regression
- Prediction results even better than simple artificial neural net
- Detection of two deposits without false positive

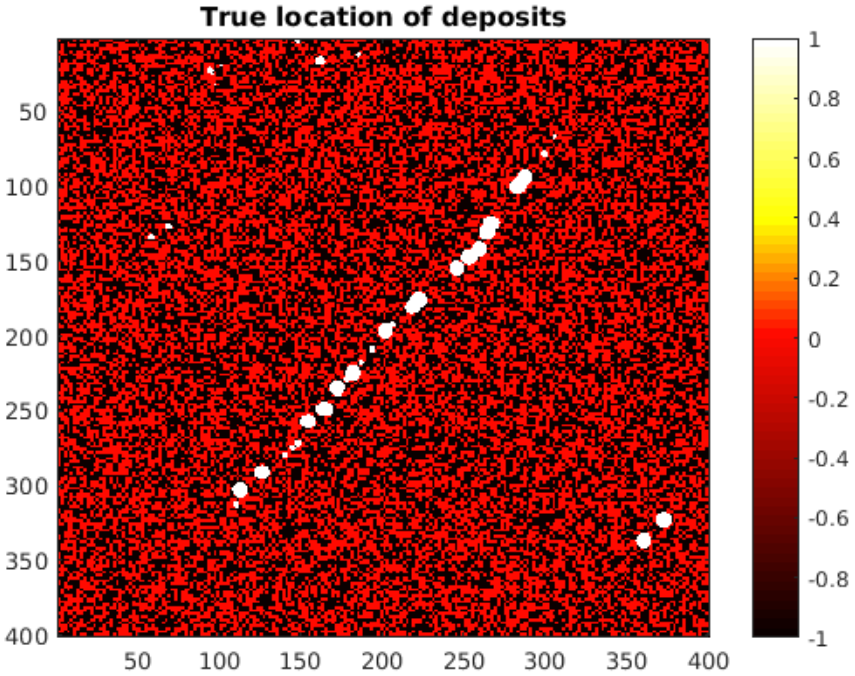
Experiments – Real Data: Search For Gold

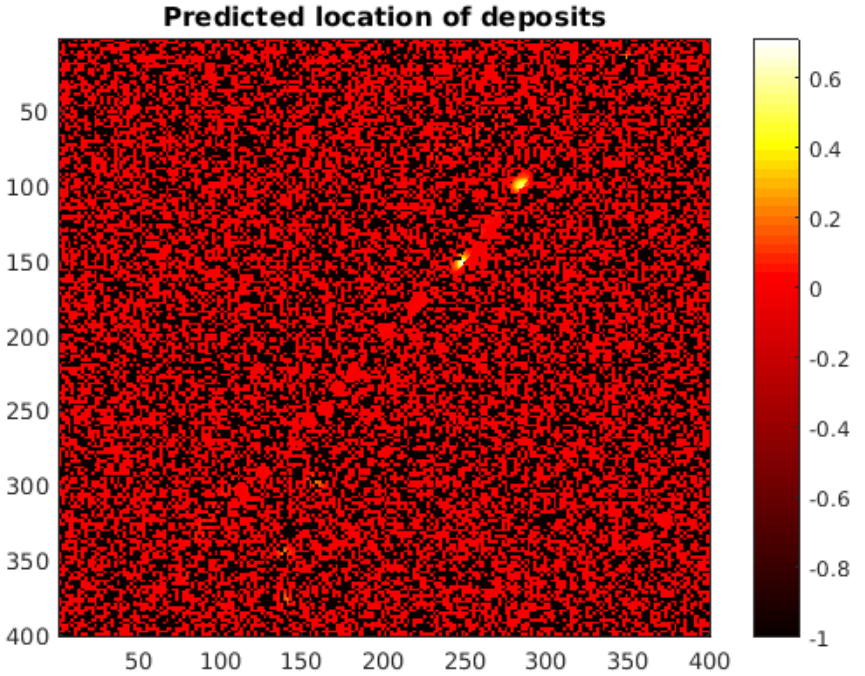


Experiments – Real Data: Search For Gold



Experiments – Real Data: Search For Gold





Conclusion

- **Model selection strategy for large data with rare events**
- **Purpose is not to give a fully automatic prediction procedure but to help guiding users what to focus on**
- **Can be used in addition to black box methods**

Thank you for your attention!

kosts@mailserver.tu-freiberg.de

The author gratefully acknowledges the support through the IAMG Travel Grant