Context / Keyword (roof)

IDA: Project 6 Caravan Insurance – Bridging the Gap between Machine Learning and Business Strategy

Keyword (backbone)

Keyword (backbone)

Chenpo HU

Reviewer: Prof. Dr. Tobias Scheffer

Tutor: Dr. Lena Jäger

Dr. Paul Prasse

Silvia Makowski

Potsdam, September 2018

Agenda

Block 1: Introduction 09:30 – 09:40

1. **Tasks and datasets**
2. **Traps that I run into**
3. **Trained machine learning process**

Block 2: Implementation and Result 09:40 – 10:10

4. **Introduction of applied machine learning methods**
5. **Implementation and result in Jupyter**
   This is the final Version of the project, in which the trained machine learning process is applied.
6. **Recommendation of good performed models**

Block 3: Discussion 10:10 – 10:20

7. **In-depth discussion**
   In-depth discussion focusing on key use-cases and customer journeys leveraging Blockchain based solutions as well as alternatives

   Aggregation of discussion results and outline of possible courses of action, including the evaluation based on the fit to business vision and analysis of strategic implications, risks and chances

Block 4: Q & A 10:20 – 10:30

8. **Exchange of thoughts**
   Do you have further improvement advices or innovative ideas?

8. **Literature**
   Literatures cited in the PPT.

Source: University Potsdam IDA, Chenpo Hu

1 Tasks and datasets

Task

Develop ***Models*** to determination of ***target customers*** of for Caravan insurance, translate into data understanding: predict as many ***class „1“*** correctly as possible

Dataset

Quantitative

- **Train dataset shape: [5822 rows x 86 columns]**
- **Output dataset shape: [4000 rows x 85 columns]**
- **Caravan Ratio: 6.36%**

Qualitative

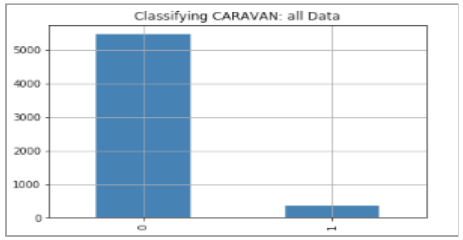
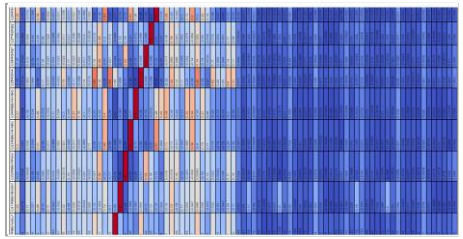
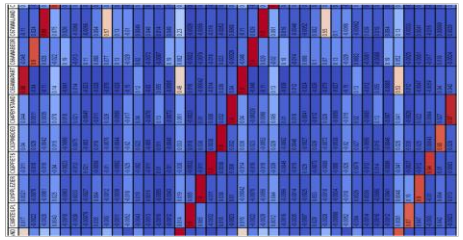
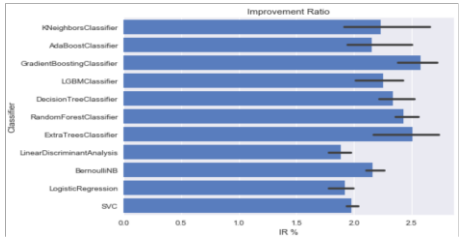
- **Demography data (1-43)**
- **Commercial data type: “Contribution of policies” (44-64)**
- **Commercial data type: “Number of policies” (65-85)**
- **Binary target class: “0” or “1” (86)**

The quality of the data seems to be too good to be true: with non null, all data are int64, but is it really so?

Source: University Potsdam IDA, Chenpo Hu

2 Traps that I run into

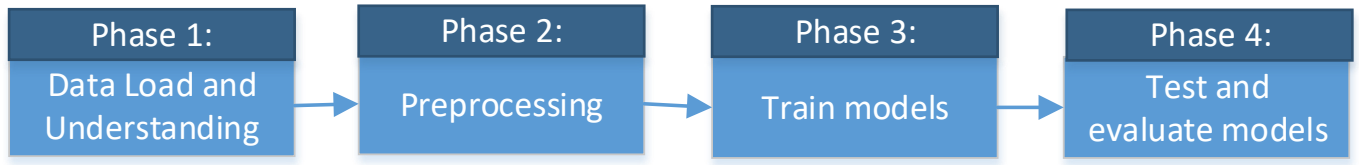
Traps !

Detours		Description	Solution
Unbalanced data		It is clear that our dataset is highly unbalanced with only **6.36% of observations** actually buying the insurance.	• Up-sample the min • Down-sample the majority class • Change your performance metric • Penalize algorithms (cost-sensitive) • Use tree-based algorithms
Into Interval coded sensitive nominal data		The **nominal data** of dataset 1 – 42 are coded into **interval** in the raw data. Some datasets in 1 - 42 should be „-1“ correlated and binary, but they are coded into scala.	Since those the **Demography data** are encrypted, we focus on the commercial data of 43-85.
Highly correlated features		Feature 44-64 and 65-85 **are strongly (at least 0,85) correlated**, since there are three red line run through the correlation matrix.	I used three kinds of features: • Feature group 1: 43-85 • Feature group 2: 43-64 • Feature group 3: Top 8 most important features
Useless standard scores		The **high accuracy score tells us nothing**, since we can say that every proband will not buy the caravan, which has the accuracy of 93.64%.	I use mainly the no. of benefit Items (TP) and the improvement ratio based on confusion matrix to evaluate the performance of the model.

Source: University Potsdam IDA, Chenpo Hu

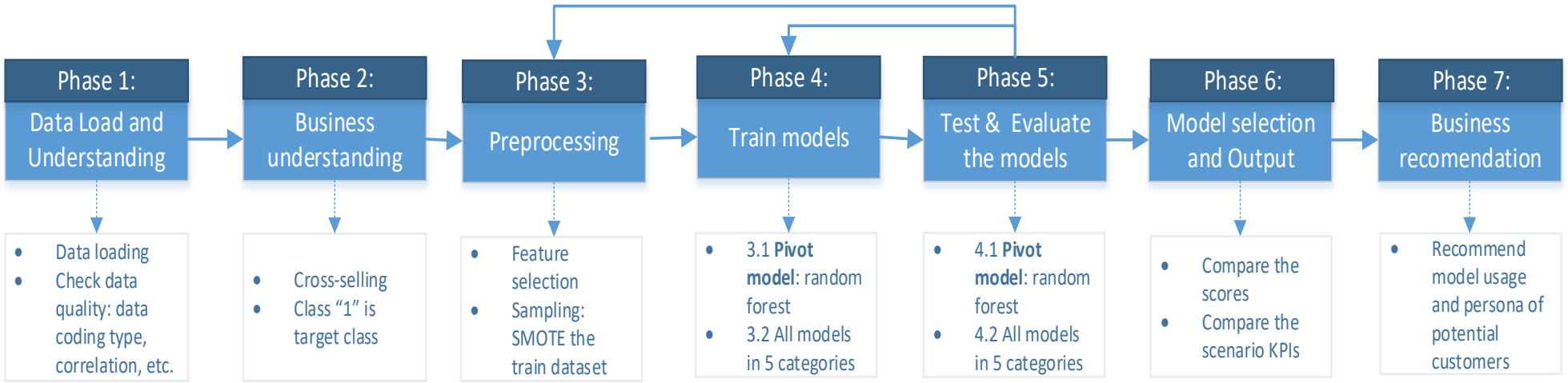
3 Trained machine learning process

Initial Process



After trap 1	Sampling the train data: SMOTE
After trap 2	Feature selection in preprocessing
After trap 3	Check statistical feature colleration in preprocessing
After trap 4	Innovate own score system and check the models in business scenarios

Trained Process



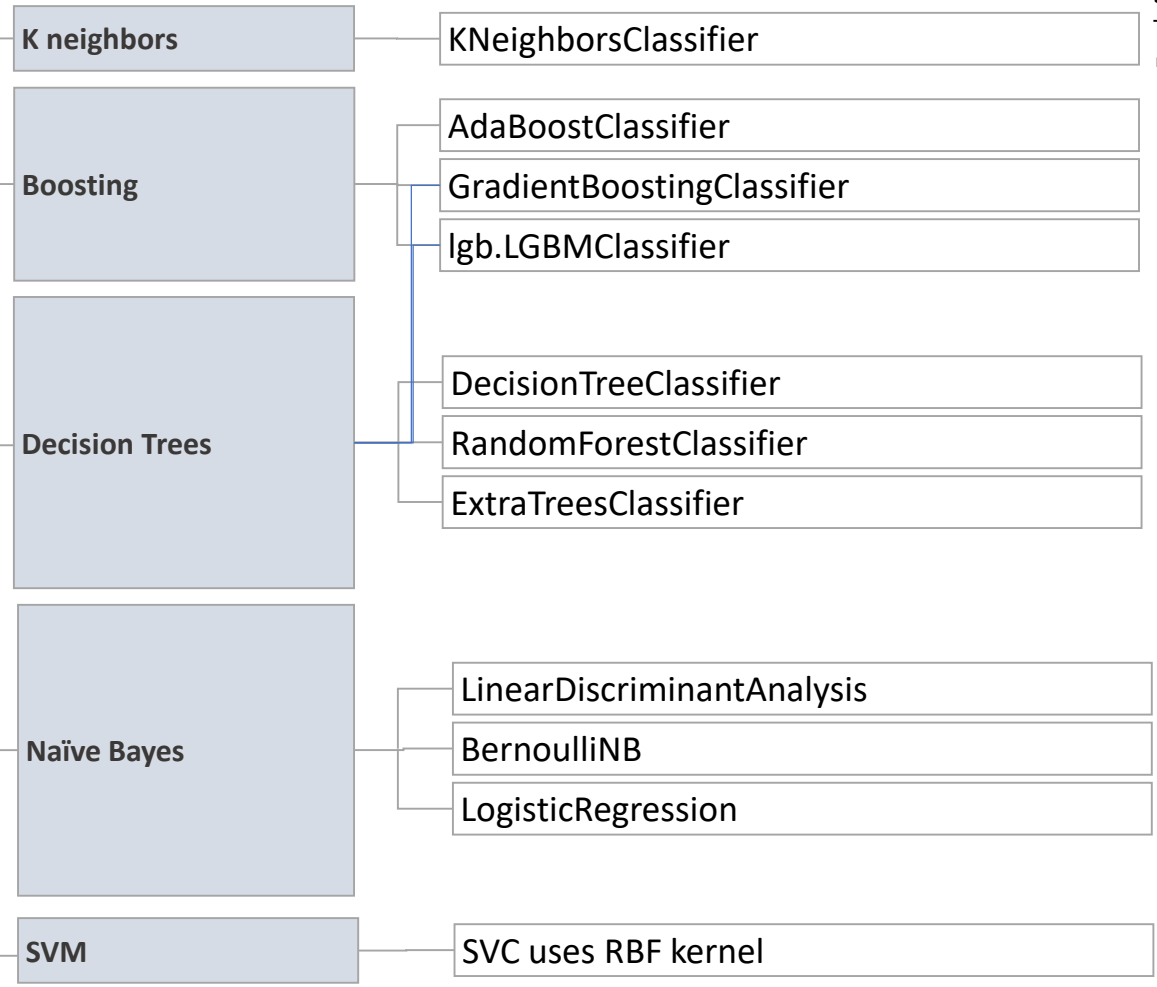
Source: University Potsdam IDA, Chenpo Hu

4 Introduction of applied machine learning methods

To select the best performed models and learn during the project more technologies, I decide to generate several categories of models instead of generating only 2-3 best performed models

5 Categories of classifiers

Overview of categorized classifiers in this project



Summary

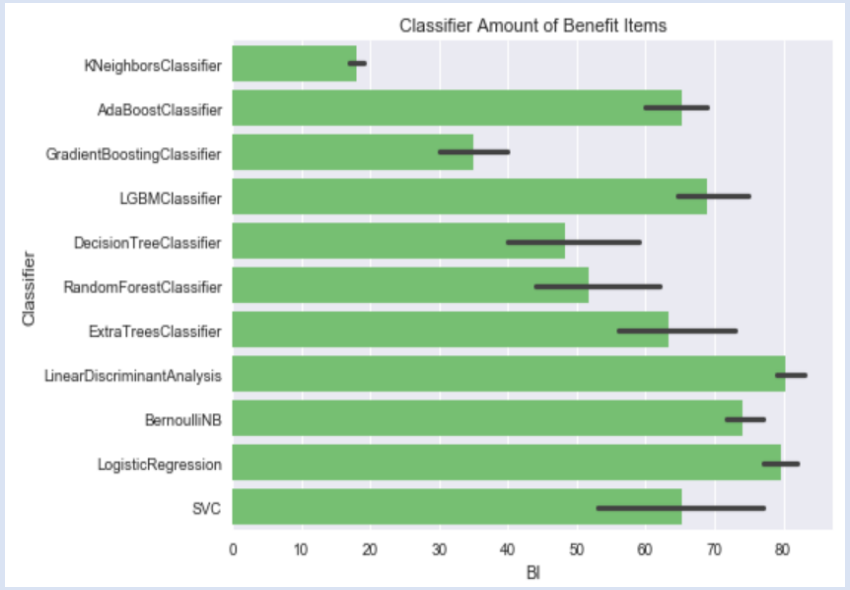
- The used classifiers can be classified **into 5 categories:**
  - **K neighbors** can be used
  - **Boosting** has a family of machine learning algorithms that convert weak learners to strong ones.
  - **Decision trees** often perform well on imbalanced datasets because their hierarchical structure allows them to learn signals from both classes.
  - **Naïve Bayes** is a frequently used in imbalanced dataset problems.
  - In handling a binary classification task, **Support Vector Machine (SVM)** is one of the methods reported to give a high accuracy in predictive modeling compared to the other techniquesuch as Discriminant Analysis s (García et al., 2015).

Source: University Potsdam IDA, Chenpo Hu

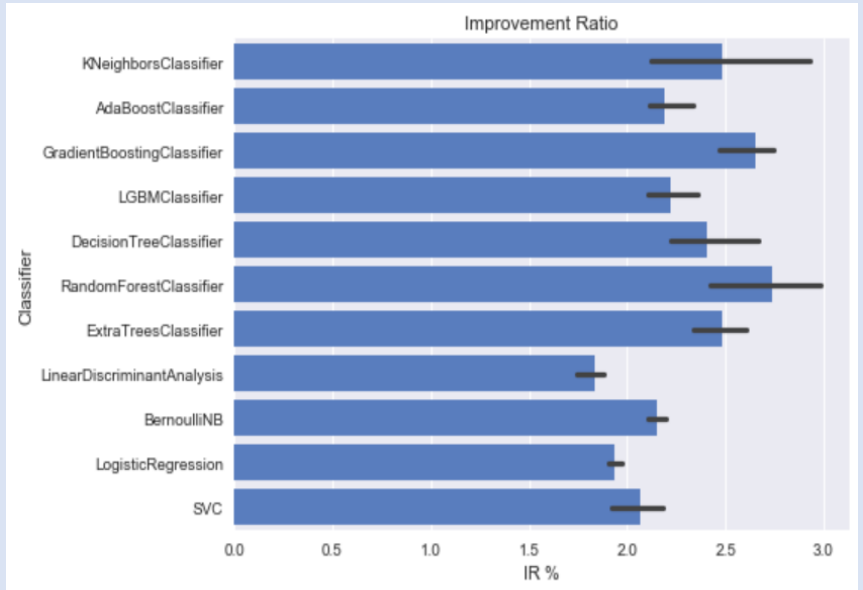
6 Recommendation of good performed models

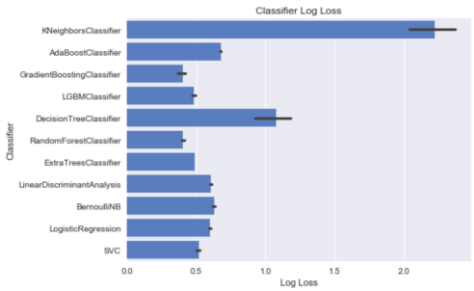
Select models

Amount of Benefit Items (correctly predicted Caravan)



Improvement Ratio





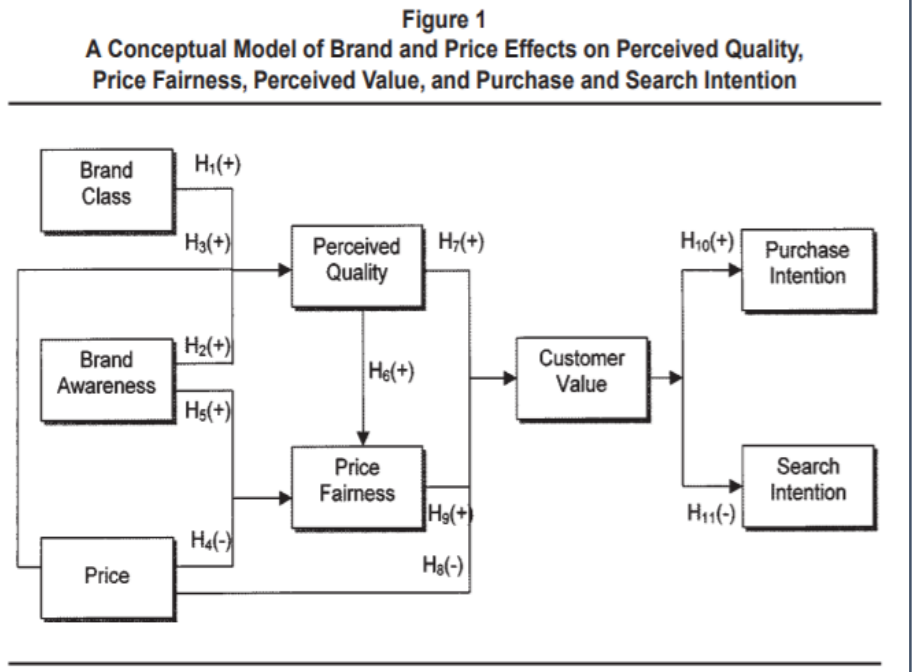
- **Improvement Ratio** = predicted true caravan rate / true caravan rate in raw data = (TP/ (TP + FP))/((TP + FN)/(TP+TN+FP+FN))
- Pros: By using the models, companies can focus more on the customers, who may like to buy caravan with less marketing cost. Companies can select the models based on their company KPIs and strategies, as realized in the scenario in the Jupyter notebook both in case to keep balance or reach the goal.
- Cons: the prediction decreased the data size, 38% - 40% (Naïve Bayes) cavavan customers may not be targeted in the marketing.

Conclusion

- **Naïve Bayes** can predict the largest amount of benefit items **(60% - 72%)** correctly, with the improvement ratio between 1.74 to 2.20. More importantly, these models has almost no overfitting problem, which may occurs in other models.
- Besides the **Naïve Bayes,** the performance of AdaBoosting, LGBMClassifer, RandomForest (but has slightly overfitting), ExtraTrees and SVC (but has slightly overfitting) are also good.
- All the models has an improvement Ratio of at least 1.74.
- The performance of RandomForestClassifeir, ExtraTreesClassifier, LogisticRegression and SVC in the models clusters using top 8 features are good and with slightly or no overfitting problem.
- The **persona of the caravan customers** are: (1) **Contribution to Car policy** with 6 (2) **high purchase power** (3) **high Contribution private third party insurance** (4) **has contribution to boat policies.**

Source: University Potsdam IDA, Chenpo Hu

7 In - depth Discussion



Source: Oh, H. (2000)

1 Maybe the false positive will become true positive in the future.

Are the false positive (customers that are predcited to be caravan, but in in turth not caravam) realy false positive?

Maybe they are the potential customers that are not digged out jet? Since this raw data is only valid in the date that it was extracted. What will happen after a few years?

2 The customers that are frequently predicted to be „1“ by different models, are most likely to be the caravan customers.

We can analyse the output excel file further and identify these customers, that are predicted to be „1“ frequently by different models.

Since then the possibility that the customer is not caravan would be the multiplication of (1-Benefitratio) of those models, which will be relativ small.

Think outside of the Box!

Source: University Potsdam IDA, Chenpo Hu

8 Q & A: Exchange of thoughts

Source: https://blitzmetrics.com/tag/big-data/

Source: University Potsdam IDA, Chenpo Hu

9 Literatures

Oh, H. (2000). The effect of brand class, brand awareness, and price on customer value and behavioral intentions. *Journal of Hospitality & Tourism Research,* 24(2), 136-162.

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (pp. 195-243). Switzerland: Springer International Publishing.

Source: University Potsdam IDA, Chenpo Hu