Automotive customer churn

ML Canvas – What needs to be considered

- Data requirements and data availability
- How predictions will be made
- Reliable and robust models
- Evaluation on monitoring consideration

Machine learning canvas

Decisions validation on value of prediction
- Evaluation metrics
- Case studies

ML tasks
- Supervised
- Unsupervised
- Regression/ Classification

Value proposition
- Model purpose statement
- Model value contribution

Data sources
- Data consolidation
- Data aggregation

Collecting data
- Surveys
- Timestamps

Making predictions
- Experimental design
- Modelling approaches

Offline evaluation
- Model complexity
- Flux in data streams

Features
- Feature extraction
- Feature selection
- Feature engineering
- Feature utility

Building models

Frequency of model re-training and feature adjustments

Live evaluation and monitoring

Dashboard for visualization of all processes that can be customized
- Utility through organisation

Louis Dorard, 2015

Value proposition

- New customer acquisition cost is **5 – 6** times higher than maintaining customers
- Retaining **5%** can increase profit **100%**

Did you know?

The US automotive industry account for 6 – 8 billion USD in yearly revenue.

ML task

- Decision trees and random forests used in predicting churn
- Suitable due to binary classification nature of these models

Collecting data

- Warranty data
- Diagnostic data
- Dealer system tracking
- General vehicle information
- Survey responses

Data sources

- Data consolidation
- Labeling
- Feature extraction
- Data balancing

Data sources

Data consolidation

Step 1: Source consolidation
- Check VIN numbers
- Match VIN numbers
- Select sufficient instances

Step 2: Specified matching
- Matching VIN and timestamps

Step 3: Tolerance matching
- Filter problematic data
- Define thresholds
- Bucket filtered data

Labelling

Choose classification boundaries

Label data accordingly

Creating classification labels
- 1 - 3: Dissatisfied
- 4 - 5: Satisfied

Feature extraction

Features used from:
- Consolidated data
- Labelled data

Feature selection is of critical importance

Feature extraction

Data balancing

Dealing with biases

Biases that we deal with here:

Class imbalance

Evaluation bias

Building models

Algorithm selection

- AdaBoost
- kNN
- SVM (RBF kernel)
- Random forests

Specify iterations

- AdaBoost (ε: $(10,20, \dots, n = 10)$)
- kNN $(1, 2, \dots, n = 10)$
- SVM $(\sum_{i=2}^{25} i = (n + 2))$
- Random forests $(\sum_{n=10}^{100} i = n)$

Training the model

- Training (0.7)
- Test(0.20)
- Validation(0.1)

Parameters and optimization

- Learning rate
- Feature partitioning
- Hyperparameter adjustments

Model considerations

- Train test split
- Preprocessing used to prepare data
- Data schematic
- Parameter adjustment

Model results

Classifier	Individual data		Aggregated data	
	Test	Validation	Test	Validation
AdaBoost	66.0	66.6	79.5	79.7
kNN	54.6	54.5	55.0	55.6
SVM (Linear)	69.6	69.6	80.3	80.2
SVM(RBF Kernel)	75.0	72.2	84.1	83.8
Radom forest	71.1	71.8	79.9	79.9

- Only used basic scoring
- Aggregate data – optimal parameter $C = 4$
- Individual data – optimal parameter $C = 16$

Evaluation

- Metric selection
- Model complexity
- Model evaluation
- Minimum description length

Evaluation - complexity

Model complexity can either:

- Improve model performance
- Decrease model performance

‘Curse of dimensionality’

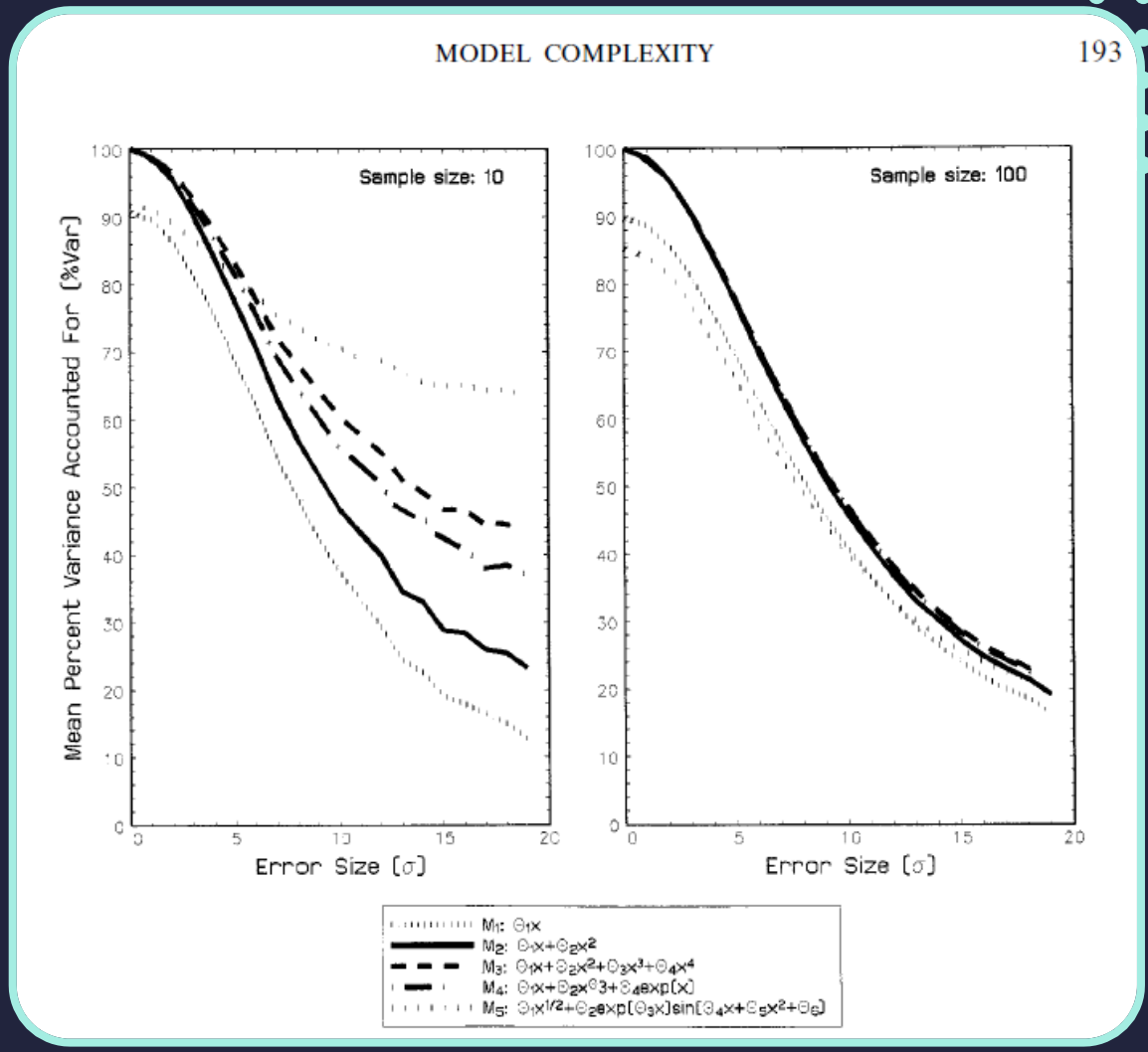


Figure 1: Model complexity versus sample sizes

Offline evaluation

- Probability distributions fluctuate
- Optimal model at any given time can fluctuate

Dashboards

Model that can adjust weights

Dashboards

- Combining model and visualization

Aggregating vestige of different steps:

Flux in features, data streams

Model performance at any given time and interaction with data

Evaluation interaction with model and data

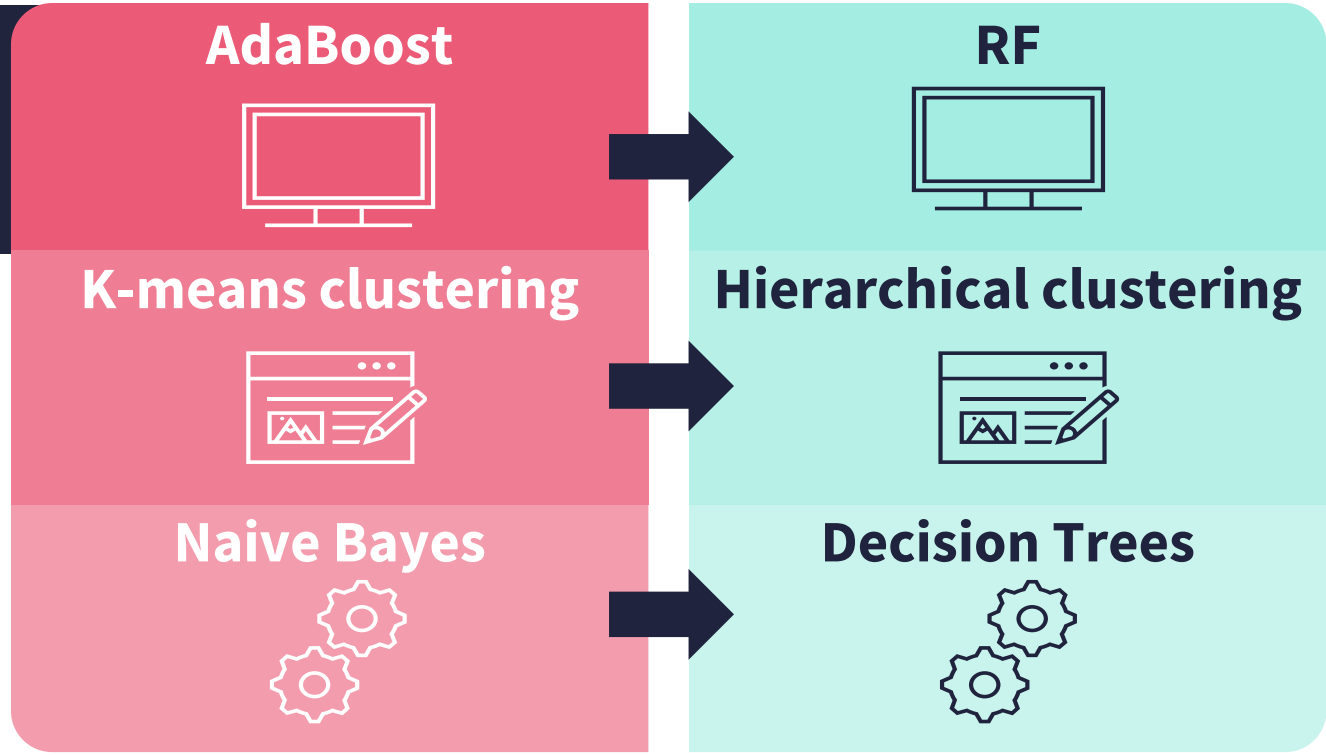
Conclusion

Considerations

Anomaly prediction

Fault diagnostics

Suggest corrective actions



Conclusion

- Reliable customer churn on simple metrics is feasible
- Larger data sources will need investigation
- More factors to consider if larger scale

List of figures

Bibliography

- In Jae Myung, The importance of complexity in model selection. Journal of Mathematical Psychology 44, 2000
- Karapınar Hasan Can, Altay Ayca, Kayakutlu G¨ulg¨un, Annals of computer science, and systems information. Churn detection and prediction in automotive supply industry. Annals of computer science and information systems, Vol 8, Pp 1349-1354 (2016, 2016)
- Yunxuan He, Ying Xiong, and Yiting Tsai. Machine learning based approaches to predict customer churn for an insurance company.2020 Systems and Information Engineering Design Symposium (SIEDS), Systems and Information Engineering Design Symposium(SIEDS), 2020, pages 1 – 6, 2020.
- Ebru Turanoglu Bekar, Per Nyqvist, and Anders Skoogh. An intelligent approach for data pre-processing and analysis in predictive maintenance with an industrial case study. Advances in Mechanical Engineering (Sage Publications Inc.), 12(5):1 – 14, 2020

Bibliography

- M. Parker. Humble Pi: A Comedy of Maths Errors. Penguin Books, Limited, 2019
- Kahneman, Daniel (2012).Thinking, fast and slow. Penguin
- Alberto Cairo [dtc] Justin Matejka [dtc] George Fitzmaurice [dtc] Lucy D'Agostino McGowan [aut]Richard Cotton [ctb] Locke Data [fnd] Steph Locke [cre, aut]. datasaurus: Datasets from thedatasaurus doze
- Louis Dorard. Machinelearningcanvas.`https: // github. com/ louisdorard/machine-learning-canvas/ projects`, 2015
- Stefan Meinzer, Ulf Jensen, Alexander Thamm, Joachim Hornegger, Björn M.Eskofier. Can machine learning techniques predict customer dissatisfaction? a feasibility study for the automotive industry. Artificial Intelligence research International Peer reviewed and Open Access Journal for Artificial Intelligence, 6:11, 2017.