ContentsGlossary

List of figures

List of tables

1 Introduction1.1 Motivation and background . . . . . . . . . . . . . . . . . . . . . . .   
1.2 Theoretical framework and focus of the study . . . . . . . . . . . . .   
1.3 Research questions and objectives . . . . . . . . . . . . . . . . . . . .   
1.4 Methodology . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
1.5 Structure of the thesis . . . . . . . . . . . . . . . . . . . . . . . . . .   
2 Machine learning: Some theoretical concepts 2.1 Data collection and preprocessing . . . . . . . . . . . . . . . . . . . .   
2.1.1 ETL . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
2.1.2 Apache spark . . . . . . . . . . . . . . . . . . . . . . . . . . .   
2.1.3 Dealing with missing data . . . . . . . . . . . . . . . . . . . .   
2.1.4 Dealing with imbalance data . . . . . . . . . . . . . . . . . .   
2.1.5 One hot encoding . . . . . . . . . . . . . . . . . . . . . . . . .   
2.1.6 Ordinal encoding . . . . . . . . . . . . . . . . . . . . . . . . .   
2.1.7 Word embedding . . . . . . . . . . . . . . . . . . . . . . . . .   
2.1.8 Data Normalization . . . . . . . . . . . . . . . . . . . . . . .   
2.1.9 Feature selection . . . . . . . . . . . . . . . . . . . . . . . . .   
2.2 Machine learning models . . . . . . . . . . . . . . . . . . . . . . . . .   
2.2.1 Logistic regression . . . . . . . . . . . . . . . . . . . . . . . .   
2.2.2 Random forest classifier . . . . . . . . . . . . . . . . . . . . .   
2.2.3 Gradient-boosted tree classifier . . . . . . . . . . . . . . . . .   
2.2.4 Decision tree classifier . . . . . . . . . . . . . . . . . . . . . .   
2.2.5 Support vector machine . . . . . . . . . . . . . . . . . . . . .   
2.2.6 Multilayer perceptron . . . . . . . . . . . . . . . . . . . . . .   
2.3 Evaluation metrics . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
2.3.1 Confusion matrix . . . . . . . . . . . . . . . . . . . . . . . . .   
2.3.2 Accuracy . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
2.3.3 NDCG@K . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
3 Related work3.1 Customer churn prediction . . . . . . . . . . . . . . . . . . . . . . . .   
3.2 Customer churn prediction in energy . . . . . . . . . . . . . . . . . .   
3.3 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4 Energy provider case study: churn prediction machine learningmodel4.1 Tools and libraries . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
4.1.1 Python . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
4.1.2 Apache Spark . . . . . . . . . . . . . . . . . . . . . . . . . . .   
4.2 Data description and understanding . . . . . . . . . . . . . . . . . .   
4.3 Data preprocessing and feature selection . . . . . . . . . . . . . . . .   
4.3.1 Handling missing data . . . . . . . . . . . . . . . . . . . . . .   
4.3.2 Dealing with categorical features . . . . . . . . . . . . . . . .   
4.3.3 Imbalanced data . . . . . . . . . . . . . . . . . . . . . . . . .   
4.3.4 Data Normalization . . . . . . . . . . . . . . . . . . . . . . .   
4.3.5 Feature selection . . . . . . . . . . . . . . . . . . . . . . . . .   
5 Models and results 5.1 Logistic regression . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
5.2 Random forest classifier . . . . . . . . . . . . . . . . . . . . . . . . .   
5.3 Gradient-boosted tree classifier . . . . . . . . . . . . . . . . . . . . .   
5.4 Decision tree classifier . . . . . . . . . . . . . . . . . . . . . . . . . .   
5.5 Support vector machine . . . . . . . . . . . . . . . . . . . . . . . . .   
5.6 Multilayer perceptron . . . . . . . . . . . . . . . . . . . . . . . . . .   
5.7 Summary and analysis of the results . . . . . . . . . . . . . . . . . .   
6 Conclusions6.1 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .   
6.2 Suggestions for future research . . . . . . . . . . . . . . . . . . . . .   
Bibliography