

Pemodelan Prediksi Customer Churn Menggunakan Teknik Augmentasi Synthetic Minority Over-sampling Technique dan Optimasi Bayesian Adaboost untuk Data Tidak Seimbang = Customer Churn Prediction Modeling Using Synthetic Minority Over-sampling Technique and Adaboost Bayesian Optimization for Data Imbalanced Handling

Doni Pradana, author

Deskripsi Lengkap: <https://lib.ui.ac.id/detail?id=9999920525922&lokasi=lokal>

Abstrak

Customer churn merupakan masalah serius di banyak sektor, termasuk sektor telekomunikasi. Pengertian customer churn adalah berhentinya penggunaan suatu layanan dan beralih ke penyedia lain atau tidak memperbarui kontrak. Untuk mengatasi risiko churn, perusahaan telekomunikasi perlu menggunakan model prediksi dengan bantuan metode machine learning. Terdapat beberapa model prediksi churn yang telah diajukan oleh para peneliti, termasuk pemilihan algoritma yang sesuai dan dataset untuk studi kasus. Pada tesis ini menggunakan dataset IBM Telco Customer Churn sebagai data pelatihan dan pengujian. Tantangan umum dalam klasifikasi adalah ketidakseimbangan data, yang dapat menyebabkan kegagalan dalam memprediksi kelas minoritas. Oleh karena itu, tesis ini menggunakan beberapa teknik augmentasi data seperti SMOTE, HAT, dan CVAE, sebagai teknik dalam menyeimbangkan data. Pembelajaran ensemble khususnya metode CART (Classification and Regression Tree) sering digunakan untuk menyelesaikan permasalahan klasifikasi dan regresi. Model Adaboost adalah algoritma pembelajaran ensemble yang menggunakan pohon keputusan sebagai dasar pembelajaran. Dalam pelatihan model Adaboost, Bayesian Optimization (BO) digunakan sebagai metode pencarian hyperparameter terbaik. Dari hasil percobaan dan pengujian yang diajukan, model Adaboost dapat memberikan nilai testing f1-score dan recall sebesar 0,661 dan 0,653 pada pelatihan dengan dataset tidak seimbang. Model Adaboost-SMOTE mempunyai nilai testing f1-score dan recall sebesar 0,646 dan 0,826. Penggunaan optimasi Bayesian Optimization pada model Adaboost-SMOTE dapat menaikkan testing f1-score dan recall menjadi 0,649 dan 0,849. Tes ANOVA dan Tukey HSD mengungkapkan variasi yang signifikan dalam hasil pelatihan dari model machine learning, dan menyoroti dampak penggunaan data seimbang dalam pelatihan model yang signifikan.

.....Customer churn is a severe problem in various sectors, including telecommunications. Customer churn refers to discontinuing the service, switching to another provider, or not renewing the contract. To deal with churn risk, telecommunication companies need to use predictive models with the help of machine learning methods. Several churn prediction models have been proposed by researchers, including the selection of suitable algorithms and data sets for case studies. In this thesis, research is conducted using the IBM Telco Customer Churn dataset. A common challenge in classification is data imbalance, which can lead to failure in predicting minority classes. Therefore, this thesis using several data augmentation techniques, such as SMOTE, HAT, and CVAE, for balancing data technique. Ensemble learning, especially the CART (Classification and Regression Tree) method, is often used to solve classification and regression problems. Adaboost is an ensemble learning algorithm that uses decision trees as the basis for learning. In the Adaboost model training, Bayesian Optimization (BO) is used to find the best hyperparameters. From the trials and tests carried out, Adaboost achieved an f1-score and recall test of 0.661 and 0.653, respectively, in training with an unbalanced dataset. The Adaboost SMOTE model achieved f1 and memory test scores of

0.646 and 0.826, respectively. Using Bayesian Optimization in the Adaboost SMOTE model increased the testing f1-score and recall scores to 0.649 and 0.849, respectively. ANOVA and Tukey HSD tests reveal significant variation in machine learning model training results and highlight the considerable impact of using balanced data in model training.