Analisis Kinerja Metode XGBoost Dan LightGBM dalam Memprediksi Klaim Asuransi Kendaraan Bermotor pada Data yang Mengandung Missing Values = Comparative Performance Analysis of LightGBM and XGBoost Methods for Predicting Motor Vehicle Insurance Claims in Datasets containing Missing Values

Rachel Aurellia Irawan, author

Abstrak

Tantangan besar dalam mengembangkan model prediktif yang baik untuk prediksi klaim asuransi kendaraan bermotor adalah adanya missing values dalam data. Berbagai algoritma pembelajaran mesin telah diteliti untuk mengatasi masalah missing values ini. XGBoost merupakan salah satu teknik Gradient Boosting Decision Tree (GBDT) yang terbukti unggul dibandingkan metode imputasi seperti K-Nearest Neighbors (KNN) dan mean imputation. Namun, XGBoost memiliki beberapa keterbatasan, seperti waktu pemrosesan yang lebih panjang dan perlunya untuk melakukan one-hot encoding pada variabel kategorikal. Keterbatasan yang dimiliki oleh metode XGBoost dapat diatasi oleh metode LightGBM. Penelitian ini bertujuan untuk menganalisis kinerja metode XGBoost dan LightGBM dalam memprediksi klaim asuransi kendaraan bermotor pada data yang mengandung missing values. Dataset yang digunakan berasal dari klaim asuransi kendaraan bermotor perusahaan Porto Seguro yang terdiri yang memiliki missing values hingga 70%. Evaluasi kinerja dilakukan menggunakan metrik Normalized Gini score dan training time. Penelitian ini membandingkan dua pendekatan dalam menangani missing values: tanpa imputasi dan dengan imputasi mean. Hasil penelitian menunjukkan bahwa metode XGBoost tanpa imputasi missing values memberikan kinerja terbaik dengan nilai Normalized Gini tertinggi sebesar 0,2735. Namun, XGBoost tanpa imputasi membutuhkan waktu training yang lebih lama, yaitu rata-rata 15,5841 detik. Metode LightGBM tanpa imputasi juga menunjukkan kinerja yang baik dengan nilai Normalized Gini sebesar 0,2559 dan waktu training yang lebih singkat dengan rata-rata 4,0521 detik. Pada data tanpa imputasi, XGBoost secara mutlak tetap menunjukkan kinerja terbaik dengan nilai Normalized Gini tertinggi baik pada data yang tidak diimputasi maupun telah diimputasi. LightGBM, meskipun memiliki Normalized Gini yang sedikit lebih rendah, namun lebih efisien dalam waktu training dengan waktu training hampir 4 kali lebih cepat dibandingkan XGBoost. XGBoost tanpa imputasi memberikan hasil prediksi yang lebih akurat. LightGBM tanpa imputasi menunjukkan efisiensi dalam waktu training dengan sedikit penurunan dalam Normalized Gini (6,88%) dibandingkan dengan XGBoost tanpa imputasi. Disimpulkan bahwa jika prioritas utama adalah kemampuan prediktif yang lebih baik, maka XGBoost tanpa imputasi adalah pilihan yang lebih baik. Namun, jika efisiensi waktu training menjadi prioritas utama, maka LightGBM tanpa imputasi dapat menjadi alternatif yang sangat baik karena mampu melakukan proses training dengan lebih cepat secara signifikan tanpa kehilangan kemampuan prediktif (dalam konteks ini Normalized Gini) yang signifikan.
.....The primary challenge in developing robust predictive models for motor vehicle insurance claims lies in the presence of missing values within the dataset. Several machine learning algorithms have been explored to address this issue, with XGBoost—a gradient-boosted decision tree (GBDT) technique—demonstrating superior performance compared to traditional imputation methods such as K-Nearest Neighbors (KNN) and mean imputation. However, XGBoost is constrained by certain limitations, including longer processing

times and the requirement for one-hot encoding of categorical variables. These limitations can be mitigated by employing the LightGBM method. This study aims to evaluate the performance of XGBoost and LightGBM in predicting motor vehicle insurance claims in datasets containing missing values. The dataset utilized in this research is sourced from Porto Seguro's motor vehicle insurance claims, which contains up to 70% missing values. The model performance is assessed using two key metrics: the Normalized Gini score and training time. The study compares two approaches to handling missing values: without imputation and with mean imputation.The findings reveal that XGBoost without imputation achieves the highest predictive performance, with a Normalized Gini score of 0.2735. However, this approach also entails a longer training time, averaging 15.5841 seconds. LightGBM without imputation, while producing a slightly lower Normalized Gini score of 0.2559, demonstrates superior efficiency, with a significantly reduced training time of 4.0521 seconds on average. In scenarios without imputation, XGBoost consistently delivers the highest predictive performance, both for non-imputed and imputed data. While LightGBM exhibits a marginally lower Normalized Gini score, it offers substantial improvements in training efficiency, with training times nearly four times faster than those of XGBoost. In conclusion, XGBoost without imputation provides the most accurate predictions, making it the preferable choice when predictive performance is the primary objective. However, when the primary concern is training time efficiency, LightGBM without imputation emerges as a strong alternative, offering a significant reduction in training time with only a modest decrease (6.88%) in predictive accuracy, as measured by the Normalized Gini score, compared to XGBoost without imputation.