INTISARI

ANALISIS MODEL CATBOOST DAN REGRESI LOGISTIK UNTUK PROBABILITAS *LAPSE* PADA POLIS ASURANSI

Oleh

Dela Agilita Septyawati

21/477899/PA/20716

Prediksi risiko *lapse* berperan penting dalam menekan potensi kerugian finansial jangka panjang bagi perusahaan asuransi. Umumnya, perusahaan asuransi mengembangkan strategi berbasis data perilaku nasabah untuk menekan risiko *lapse*. Model regresi logistik populer digunakan karena sederhana dan mudah diinterpretasikan. Namun, dengan tersedianya data yang lebih terperinci, model regresi logistik yang mengasumsikan hubungan linear antara prediktor dan *log-odds* target tidak mampu menangkap interaksi yang kompleks dan non-linear dalam data. Penelitian ini mengevaluasi performa CatBoost, sebuah algoritma *ensemble gradient boosting*, untuk memprediksi polis *lapse*. CatBoost dapat langsung menangani fitur kategorikal tanpa *encoding* tambahan, sekaligus mampu menangkap interaksi non-linear. Menggunakan dua data dengan karakteristik berbeda, berdasarkan domain, proporsi target, dan kumpulan fitur, penelitian ini menunjukkan bahwa CatBoost mengungguli *recall* hingga 66% dibandingkan regresi logistik dengan *recall* 0 dan 20% pada data asli yang tidak seimbang. Dengan performa yang lebih baik, CatBoost memberikan informasi yang lebih akurat bagi perusahaan dalam menyusun strategi manajemen risiko, sehingga retensi nasabah dapat ditingkatkan.

ABSTRACT

ANALYZING CATBOOST AND LOGISTIC REGRESSION MODELS FOR LAPSE PREDICTION ON INSURANCE POLICIES

By

Dela Agilita Septyawati

21/477899/PA/20716

Policy lapse prediction plays a critical role in mitigating long-term financial losses for insurance companies. Typically, insurance companies develop targeted strategies to reduce the risk of lapse by analyzing customer behavioral data. To this end, logistic regression is usually picked as the model of choice. However, with more granular behavioral data becoming available, the linear nature of logistic regression model fails to capture the more complex and non-linear interactions in the dataset. In this thesis, we study the performance of CatBoost, an ensemble gradient boosting algorithm for policy lapse prediction. CatBoost is capable of natively handling categorical features without requiring additional encoding as well as non-linear interaction. We evaluate CatBoost on two datasets with distinct characteristics (i.e., different domains, target distributions, and feature sets). Our analysis shows that CatBoost substantially outperforms logistic regression in recall, achieving 66% compared to 0 and 20% on the original imbalanced data. With higher recall, CatBoost offers potential for a more informed risk management strategy, which in turn improves the insurer's customer retention.