**PENDAHULUAN**

Fraud detection adalah proses mengidentifikasi tindakan pengguna dalam suatu skenario dikategorikan sebagai tindakan penipuan atau bukan. Dalam konteks kompetisi ini, tindakan penipuan didefinisikan sebagai pengguna platform yang telah meminjam produk keuangan tetapi tercatat belum melakukan pembayaran sampai tenggat waktu yang telah ditentukan.

Machine learning memainkan peran penting dalam mengatasi permasalahan *fraud detection*. Salah satu model machine learning yang dapat digunakan untuk menyelesaikan permasalahan ini adalah *logistic regression*. *Logistic regression* adalah model pembelajaran mesin untuk klasifikasi. Model ini digunakan untuk memprediksi kemungkinan variabel dependen kategorikal. (Kumar *et al*. 2020).

Model *logistic regression* dipilih karena regresi logistik mudah diterapkan dan hasilnya mudah diinterpretasikan. Hal ini penting dalam deteksi penipuan, dimana memahami hubungan antara prediktor dan kemungkinan penipuan bisa menjadi sangat penting. Selain itu, regresi logistik menampilkan skor presisi rata-rata yang lebih baik dari beberapa algoritma yang telah digunakan.

**ANALISIS DAN PEMBAHASAN HASIL**

Pada pengembangan model ini, Kami dihadapkan pada beberapa tantangan untuk mengidentifikasi fraud pada pengguna *platform* f*intech* yaitu:

1. Informasi dari data yang sangat besar. Dataset yang diberikan memiliki hampir satu juta observasi. Tidak hanya jumlah observasi yang besar, dataset juga memiliki fitur yang cukup banyak
2. Adanya duplikasi dari dataset namun dengan pelabelan yang berbeda. Hal ini menyebabkan sulitnya pengambilan keputusan untuk mempertahankan data atau tidak. Hampir 400 ribu data terdeteksi duplikat, yang apabila dilakukan penghapusan dapat menyebabkan hilangnya hingga 90 persen data dengan label 1 atau terdeteksi fraud.
3. Variabel antar dataset yang berkorelasi sangat tinggi, sebagai contoh pc4 dan pc9 memiliki nilai korelasi sempurna yaitu 1. Hal dapat mengganggu model, namun cukup sulit untuk memutuskan penghapusan variabel tersebut dikarenakan dilakukannya pengaburan pada setiap variabel sehingga sulit diidentifikasi.
4. Informasi yang tidak seimbang, yaitu sebagian besar transaksi 98.74% bukan merupakan penipuan, sehingga sulit untuk mengidentifikasi aktivitas penipuan.
5. Adanya standarisasi atau normalisasi data yang dilakukan pada data sebelumnya mengakibatkan data tersebut lebih abstrak dan menyulitkan dalam memahami pola data.

EDA

Informasi dari dataset memiliki ukuran yang sangat besar di mana dataset yang diberikan memiliki hampir satu juta observasi. Tidak hanya jumlah observasi yang besar, namun juga memiliki fitur yang banyak. Sayangnya, informasi yang diberikan tidak seimbang, yaitu sebagian besar transaksi yaitu sebesar 98.74% bukan merupakan penipuan (Non Fraud) sehingga sulit untuk mengidentifikasi aktivitas penipuan. Seluruh dataset yang digunakan untuk model pelatihan tidak menunjukkan adanya *missing values*. Selain itu, adanya hampir 400 ribu data duplikasi observasi pada seluruh nilai fitur tetapi dengan pelabelan yang berbeda. Hal ini menyebabkan sulitnya pengambilan keputusan untuk tetap atau tidaknya mempertahankan data. Dengan dataset asli yang sudah dinormalisasi, maka variabel-variabel (pc0-pc16) tampaknya memiliki rentang yang sangat kecil, dengan nilai median mendekati nol. Kotak-kotaknya sangat kecil yang terlihat pada box plot yang dihasilkan menunjukkan variasi minimal di mana nilai-nilai variabel sangat dekat satu sama lain dan distribusinya sempit. Artinya, tidak banyak perbedaan antara nilai variabel dan sebagian besar nilai berkumpul di sekitar pusat distribusi tanpa banyak outlier atau nilai ekstrim. Hal ini mengakibatkan data tersebut lebih abstrak dan menyulitkan dalam memahami pola data. Pada analisis multicollinearity, didapatkan hasil yang menunjukkan bahwa antar variabel memiliki multicollinearity yang tinggi. Sebagai contoh, variabel pc4 dengan pc9, pc8 dengan pc6, memiliki nilai korelasi sempurna yaitu 1. Hal ini dapat mengganggu model karena dapat menyebabkan terjadinya kesalahan dalam prediksi, yaitu *overfitting*. Multicollinearity memicu masalah *overfitting* karena model memberikan bobot berlebihan pada variabel yang sangat berkorelasi. Hal ini mengakibatkan model menjadi lebih rumit dan kurang mampu membuat prediksi pada data yang belum pernah ada sebelumnya yang mengakibatkan terjadinya penurunan skor akurasi prediksi. Selain itu, variabel antar dataset yang berkorelasi sangat tinggi dapat pula mengakibatkan redudansi variabel di mana antar variabel yang berkorelasi tinggi membawa informasi yang sama. Hal ini menambah kompleksitas model tanpa meningkatkan kemampuan prediktifnya. Redudansi tidak hanya membuat model lebih sulit memprediksi, tetapi juga memperlambat proses pelatihan dan dapat memperburuk performa model. Namun, cukup sulit bagi kami untuk memutuskan menghapus variabel-variabel tersebut karena variabel tersebut sengaja dikaburkan untuk kepentingan anonimitas.

Preprocessing

Pada proses ini kami melakukan pemeriksaan pada data null (*missing value*) dan duplikat. Untuk label 1 hasilnya sebanyak 7109 dan untuk label 0 ada sebanyak 303395

Kami memutuskan untuk menghapus duplicate pada data berlabel 0 karena berdasarkan EDA, diperoleh fakta bahwa terdapat ketidakseimbangan pada data label *fraud* dan *nonfraud* yang cukup timpang di mana data dengan label 1 ada sebanyak 10857 dan data dengan label 0 berjumlah 78x lipat dibandingkan data berlabel 1, yakni sebanyak 847042. Sebagai bagian dari solusi untuk mengatasi *imbalanced data*, kami memutuskan untuk menghapus data dengan label 0 (*fraud*).

Ditemukan pula bahwa pola pc0 = 0 dan pc1 - pc16 = -1 kecuali pc10 = 0.0, serta pc0 = 1 dan pc1 - pc16 = -1 kecuali pc10 = 0.0 sering muncul dalam dataset non\_borrower\_user. Mengingat bahwa dataset non\_borrower\_user tidak digunakan dalam pelatihan model, hal ini memperkuat alasan untuk menghapus seluruh data yang memiliki pola fitur identitas tersebut.

MODELING

Model regresi logistikdipilih karena mudah diterapkan dan hasilnya mudah diinterpretasikan. Hal ini penting dalam deteksi penipuan, dimana memahami hubungan antara prediktor dan kemungkinan penipuan bisa menjadi sangat penting. Selain itu, regresi logistik dengan regularisasi (L1 atau L2) dapat membantu mengatasi multikolinearitas dengan memberikan penalti pada koefisien yang besar, sehingga mengurangi dampaknya dan berpotensi menghasilkan performa dan stabilitas model yang lebih baik. Regresi logistik juga memberikan perkiraan probabilitas, yang bisa sangat berguna dalam deteksi penipuan untuk menetapkan ambang batas dan membuat keputusan berbasis risiko.

Pada proses prediksi dataset yang diberikan dilihat pula bahwa dataset tersebut beberapa nilai yang tidak diketahui labelnya atau memiliki label ganda apabila dilihat pada data *train*. Hal ini memberikan keuntungan apabila menggunakan LR, dimana LR sangat efisien untuk mengklasifikasikan catatan yang tidak diketahui. Selain itu, berdasarkan beberapa percobaan yang telah dilakukan, jika dibandingkan dengan algoritma LightGBM, FA-CNN, RLS, model LR memberikan skor presisi rata-rata yang lebih baik. LightGBM memberikan *AP score* , FA-CNN , RLS 0.6774, dan LR 0.8138. Hal ini menunjukkan bahwa model LR memberikan nilai *average precision score* yang lebih baik.

**KESIMPULAN**

Dari analisis yang telah dilakukan, dapat disimpulkan bahwa …

**DAFTAR PUSTAKA (pake format IPB)**

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4135514>

Ali A, Razak SA, Othman SH, Eisa TAE, Al-Daqhm A, Nasser M, Elhassan T, Elshafie H, Saif A. 2022. *Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review*. *Applied Sciences*. 12(9): 9637. <https://doi.org/10.3390/app12199637>.

James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning: with Applications in R.* New York (NY): Springer

Kumar Y, S Sameeka, Payal R. 2020. Comparative analysis for fraud detection using logistic regression, random forest, and support vector machine. *SSRN Electric Journal*. 7(4):728. http://dx.doi.org/10.2139/ssrn.3751339.

Trivedi NK, Simaiya S, Lilhore UK, Sharma SK. 2020. *An Efficient Credit Card Detection Model Based on Machine Learning Method*. *International Journal of Advanced Science and Technology*. 29 (5): 3414 - 3424.