Easy Report

Credit Risk Score Project

1. Background

credit risk score adalah metode pengendalian risiko yang umum di industri keuangan. Ini menggunakan informasi pribadi dan data yang dikirimkan oleh pemohon kartu kredit untuk memprediksi kemungkinan gagal bayar dan pinjaman kartu kredit di masa mendatang. Bank dapat memutuskan apakah akan menerbitkan kartu kredit kepada pemohon. Credit Scoret dapat mengukur besarnya risiko secara objektif.

Secara umum, credit risk score didasarkan pada data historis. Setelah menghadapi fluktuasi ekonomi yang besar. Model sebelumnya mungkin kehilangan kekuatan prediktif aslinya.

Dalam project ini saya mengambil kasus credit risk score, jadi di dalam perusahaan penyedia kredit, tentunya pihak perusahaan ingin memberikan credit kepada pihak-pihak yang capable untuk memenuhi hak-hak dan kewajiban-kewajiban yang telah disetujui. Terjadi banyak kasus dimana pihak-pihak tertentu lalai dan tidak dapat memenuhi kewajiban-kewajiban dengan alasan tertentu. Oleh karena itu, perlu kehati-hatian dari pihak penyedia kredit dalam memberikan credit.

Untuk mengatasi masalah tersebut, motivasi saya adalah ingin membuat suatu model dimana kepada siapa saja penyedia credit bisa/ dapat memberikan credit secara aman, tanpa takut terjadi telat bayar atau malah tidak dibayar. Dengan menggunakan fitur-fitur dan metode tertentu saya berharap dapat membedakan mana user yang capable diberikan credit dan mana yang tidak.

Input permasalah ini adalah beberapa fitur seperti pendapatan, jumlah anak dan lain sebagainya, kemudian saya menggunakan Random Forest untuk memprediksi siapa saja yang capable untuk diberikan credit dan siapa yang tidak.

2. Related Work

a. https://www.researchgate.net/publication/5144412_Credit_Risk_Assessment_Using_Statistical_and_Machine_Learning_Basic_Methodology_and_Risk_Modeling_Applications
b. https://www.mdpi.com/2227-9091/6/2/38
c. https://www.sciencedirect.com/science/article/pii/S1877050920315830

Project tersbut menggunakan metode Decision Tree, KNN, dan Neural Network

3. Dataset & Features

Untuk preprocessing saya melakukan pengecekan terhadap duplikasi, redundancy, inkonsistensi, dan missing value. Tidak ada duplikasi dan inkonsistensi data dalam dataset tersebut. Saya membuang kolom FLAG_MOBIL karena kolom tersebut hanya berisi 1 macam informasi, jadi tidak perlu digunakan untuk permodelan.

Pada dataset tersebut terdapat kolom categorical yang memiliki missing value, untuk itu saya melakukan imputasi dengan mengisi kekosongan data dengan string "KOSONG"

Targetnya adalah kolom STATUS berisi X,C,0,1,2,3,4,5. Saya membagi kolom target menjadi 2 yaitu X dan C menjadi kolom yang aman dan sisanya termasuk mencurigakan. kemudian saya membuat kolom bad_flag yang berisi binary, dimana user yang termasuk kedalam kelompok mencurigakan akan mendapat flag 1 atau yes sedangkan yang aman akan mendapat flag 0 atau no.

Saya mengecek proporsi data melalui fungsi value_counts, apakah terjadi inkonsistensi pada data bisa dilihat dari sini. Kemudian Saya membagi data training dan test dengan rasio 80:20 dengan nilai random_state adalah 2, Setelah semua data sudah bersih, tidak ada missing value, duplikasi, dan inkonsistensi kemudian saya merubah data kategorikal ataupun string menjadi integer atau float, hal ini wajib dilakukan karena untuk melakukan permodelan data yang diproses adalah numerikal, tidak bisa kategorikal. Saya menggunakan One Hot Encoding untuk mengekstrak data tersebut.

Setelah semua data yang dibutuhkan sudah berupa numerikal maka saya kemudian melakukan standardisasi, hal tersebut bertujuan untuk menjadikan rentang data ke dalam skala yang sama, misalkan pada kolom jumlah anak maksimal hanya teris nilai 5 sedangkan pada kolom pendapatan nilainya bisa sampai ratusan ribu, untuk itu harus dijadikan ke dalam skala yang sama agar bisa masuk ke tahap permodelan. Pada proses ini saya menggunakan fitur dari sklearn yaitu StandardScaler.

Meskipun dalam pembuatan model decision tree tidak dibutuhkan standardisasi, karena ide dasar dari decision tree adalah menentukan Treshold, jadi tidak memikirkan jarak antar data. Proses Standardisasi ini dilakukan hanya pada data Training saja, tidak pada data Test, karaena untuk menghindari data leakage.

Pada proses ini, karena saya melakukan dengan metode klasifikasi maka untuk menentukan baseline, saya menggunakan majority vote atau proporsi data target yaitu kolom bad_flag. Dapat kita lihat bahwa baseline akurasinya adalah 61%.

Sumber dataset https://www.kaggle.com/datasets/rikdifos/credit-card-approval-prediction

4. Methods

Random forest Adalah salah satu cara untuk membuat suatu kumpulan dari descision tree, dengan metode Random Forest, saya bisa membuat model yang lebih uncorrelated.

Jika ingin memprediksi apakah seseorang termasuk bad_flag atau tidak berdasarkan fitur-fitur yang ada seperti Pendapatan, Jumlah anak, Tipe rumah, Jenis pekerjaan dan lain sebagaianya.

Dalam metode ini saya ingin menambahkan fitur randomisasi pada setiap fitur-fitur yang ada untuk permodelan, jadi randomisasi dilakukan tidak hanya pada tahap data, melainkan pada tahap fiturnya. Output dari proses ini adalah banyaknya model yang prediksi, dimana hasil akhrinya adalah agregasi dari seluruh hasil prediksi tersebut.

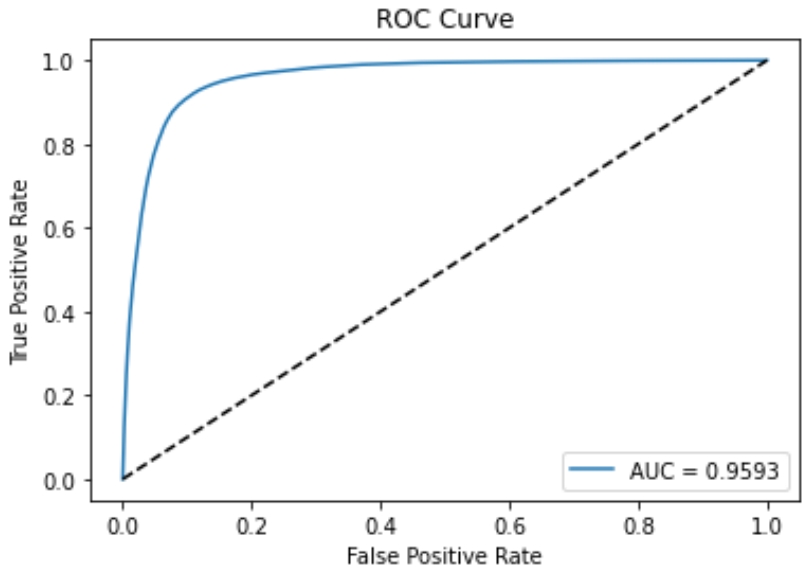
Dengan menggunakan Random Forest, masalah overfitting dari kedalam tree otomatis sudah terselesaikan dengan mengurangi variansi modelnya dan juga bisa mengkoreksi satu sama lain. Dengan menggunakan metode Random Forest, fitur-fitur yang diprediksi akan semakin spesifik, lalu model akan lebih uncorrelated, sehingga performa bisa lebih tinggi lagi.

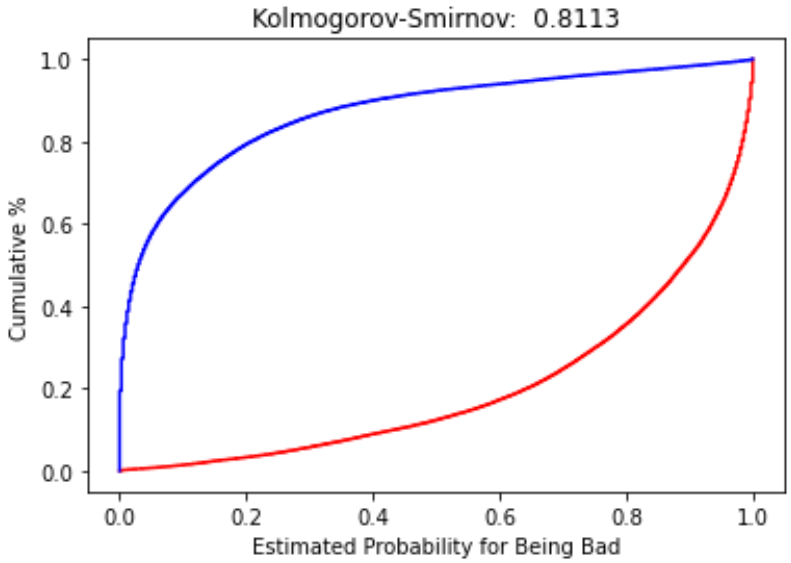
5. Experiment and Result

Dalam Random Forest, HyperParameter yang digunakan adalah jumlah model yang dibuat(n_estimator), max_depth, dan jumlah fitur. Jumlah model yang dibuat (Tree) mempengaruhi classification error, semakin banyak n_estimator maka semakin stabil errornya, jadi hyperparameter ini adalah salah satu yang faktor penting dalam metode Random Forest.

Hasil score acccuracy dari model kita adalah 0.9 dimana sudah berhasil mengalahkan model. Untuk mengukur performa model, dua metrik yang umum digunakan dalam dunia credit risk adalah AUC dan Kolmogorov-Smirnov, untuk itu saya menggunakan dua metrik ini dalam project saya.

Model yang dibangun menghasilkan performa AUC = 0.9593 dan KS = 0.8113, Pada dunia credit risk modelling, umumnya AUC di atas 0.7 dan KS di atas 0.3 sudah termasuk performa yang baik. Model yang baik adalah model yang nilai AUC-nya mendekati angka 1, kita bisa melihat bahwa luas daerah dibawah kurva AUC cukup besar/ tinggi, hal tersebut menunjukkan bahwa model kita memiliki classifier yang bagus, regardless apapun tresholdnya.





6. Conclusion

Menurut saya kenapa saya menggunakan Random Forest, karena dengan metode tersebut saya bisa membuat model lebih uncorrelated, dengan Random Forest juga dapat mengatasi masalah overfitting dari kedalaman tree yang otomatis sudah terselesaikan dengan mengurangi variansi modelnya serta juga bisa mengkoreksi satu sama lain. Kesimpulannya model ini cukup bisa menjawab permasalahan yang terjadi

References:

1. https://towardsdatascience.com/evaluating-classification-models-with-kolmogorov-smirnov-ks-test-e211025f5573
2. https://qualityamerica.com/LSS-Knowledge-Center/statisticalinference/k-s_statistic.php