Loan Prediction Based on Customer Behavior

By Group 2: Pearson Explorer

Contents

BACKGROUND

What is the problem?

Organisasi ingin memprediksi pelanggan yang berpotensi mangkir untuk produk pinjaman konsumen berdasarkan data perilaku pelanggan historis yang mereka amati. Dengan demikian, mereka dapat mengidentifikasi pelanggan baru yang lebih berisiko dan mengambil tindakan proaktif untuk mengelola risiko default.

Supporting Facts:

- Menurut laporan industri keuangan, risiko default pelanggan dapat menyebabkan kerugian besar bagi perusahaan, terutama jika tidak dideteksi secara dini (source).
- Penelitian menunjukkan bahwa menggunakan teknik analisis data dan machine learning dapat membantu mengidentifikasi pola dan tren yang berkaitan dengan risiko default pelanggan (source).

Problem Statements

- ✔ Menentukan apakah seorang pelanggan berisiko default pada pinjaman berdasarkan perilaku dan karakteristiknya.
- ✔ Membuat model prediksi untuk mengidentifikasi pelanggan yang berpotensi gagal membayar pinjaman..
- ✔ Menganalisis faktor-faktor yang mempengaruhi risiko default pada pinjaman.
- ✔ Menentukan apakah ada korelasi antara variabel seperti pendapatan, usia, atau pengalaman dengan risiko default pada pinjaman.

What do we do?

Sebagai tim Data Scientist, Kami menggunakan pengetahuan dan keterampilan analisis data untuk memahami tren dan pola dalam dataset pinjaman. Bertanggung jawab untuk mengembangkan model prediksi yang dapat mengidentifikasi pelanggan yang berisiko default pada pinjaman berdasarkan fitur-fitur seperti pendapatan, usia, pengalaman, dan karakteristik lainnya. Selain itu, Kami juga bertugas untuk menganalisis faktor-faktor yang mempengaruhi risiko default dan menghasilkan wawasan yang berguna bagi tim manajemen atau tim pengambil keputusan dalam mengelola portofolio pinjaman. Dengan menggunakan pendekatan analisis data dan teknik machine learning, Kami berperan dalam membantu perusahaan atau institusi keuangan mengoptimalkan keputusan kredit dan mengurangi risiko gagal bayar dalam portofolio pinjamannya.

Goal & Objectiveness

GOAL

Mengurangi risiko gagal bayar dalam portofolio pinjaman dan meningkatkan kualitas keputusan kredit dengan menggunakan analisis data dan model prediksi.

01

Membangun model prediksi yang dapat mengidentifikasi pelanggan yang berpotensi default pada pinjaman dengan akurasi tinggi

02

Meningkatkan pemahaman tentang faktor-faktor yang mempengaruhi risiko default melalui analisis data yang mendalam..

03

Mengoptimalkan keputusan kredit dengan menggunakan model prediksi untuk menilai risiko default pada setiap aplikasi pinjaman.

04

Meningkatkan Repayment Rate pada portofolio pinjaman pelanggan.

Goal

Business Metrics

Repayment Rate

Persentase dari total pinjaman konsumen yang berhasil bayar (Non Default) dalam portofolio pinjaman. Metric ini membantu dalam melacak apakah jumlah pelanggan yang mangkir berkurang sesuai dengan tujuan untuk mengurangi risiko gagal bayar.

Tingkat Akurasi Model Prediksi

Persentase dari prediksi model yang benar dari total prediksi yang dilakukan oleh model. Metric ini membantu dalam mengevaluasi seberapa baik model dapat mengidentifikasi pelanggan yang berpotensi mangkir dengan akurat.

Presisi dan Recall

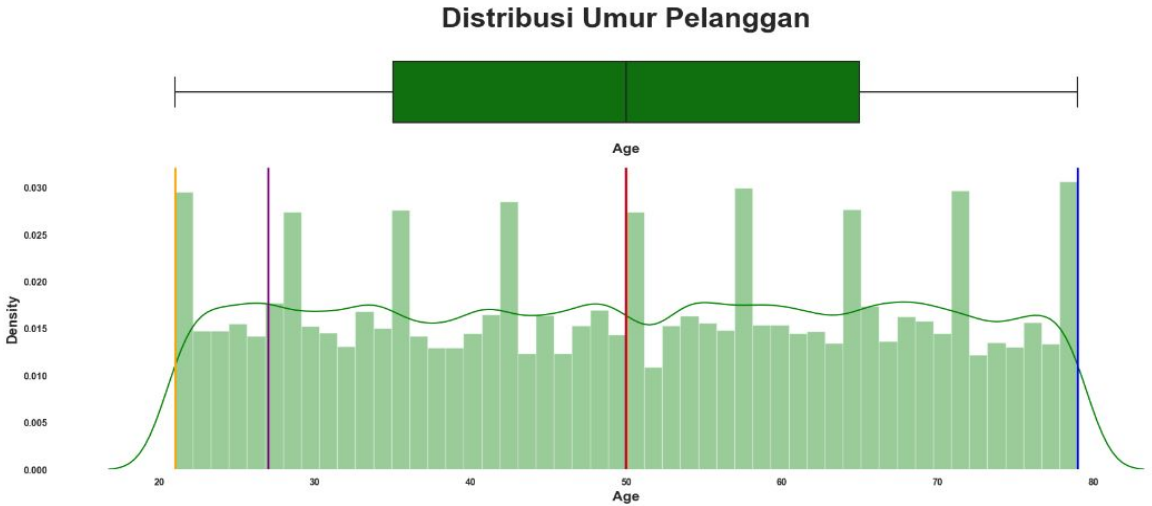
Persentase presisi (positif prediksi yang benar) dan recall (kasus positif yang diidentifikasi) dari model prediksi. Metric ini membantu dalam mengevaluasi seberapa baik model dapat mengelola risiko mangkir dengan mengoptimalkan jumlah pelanggan yang benar-benar berisiko mangkir dan meminimalkan jumlah pelanggan yang tidak terdeteksi sebagai berisiko mangkir.

1. EDA

Exploratory Data Analyst

UNIVARIATE ANALYSIS

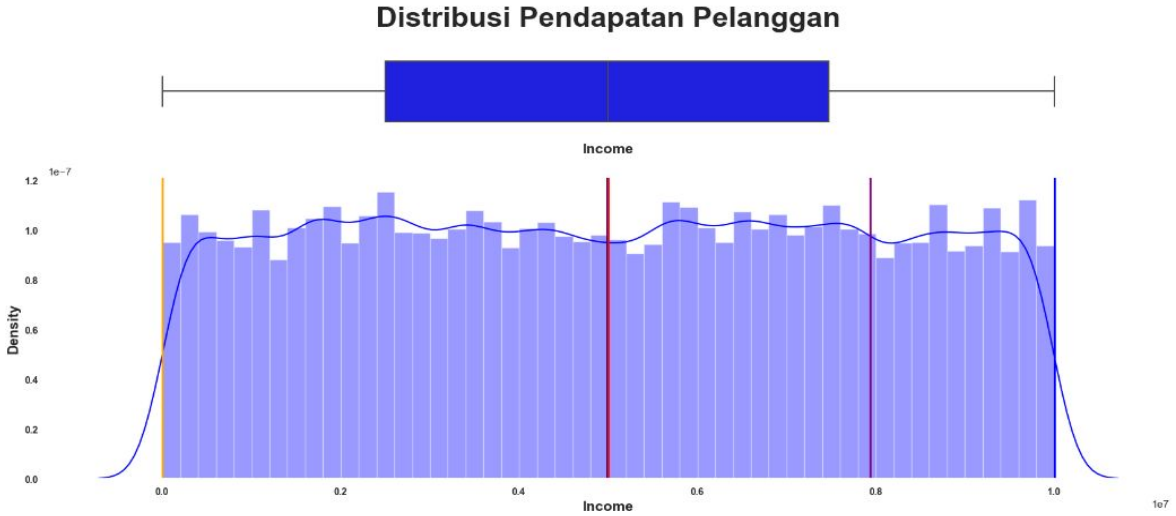
mean=50.0
median=50.0
max=79
min=21
mode=27



AGE Column

UNIVARIATE ANALYSIS

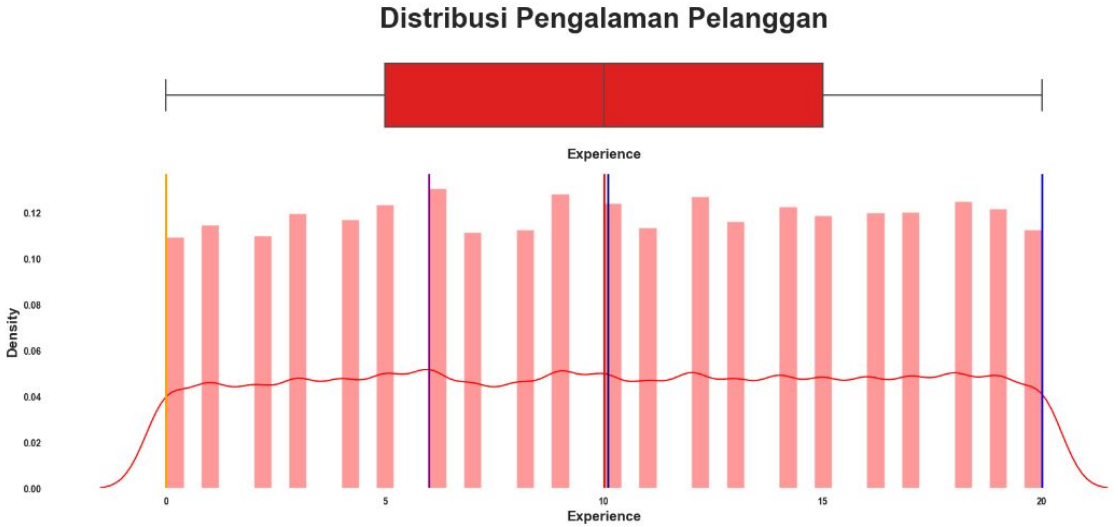
mean=4997116.7
median=5000694.5
max=9999938
min=10310
mode=7936020



INCOME Column

UNIVARIATE ANALYSIS

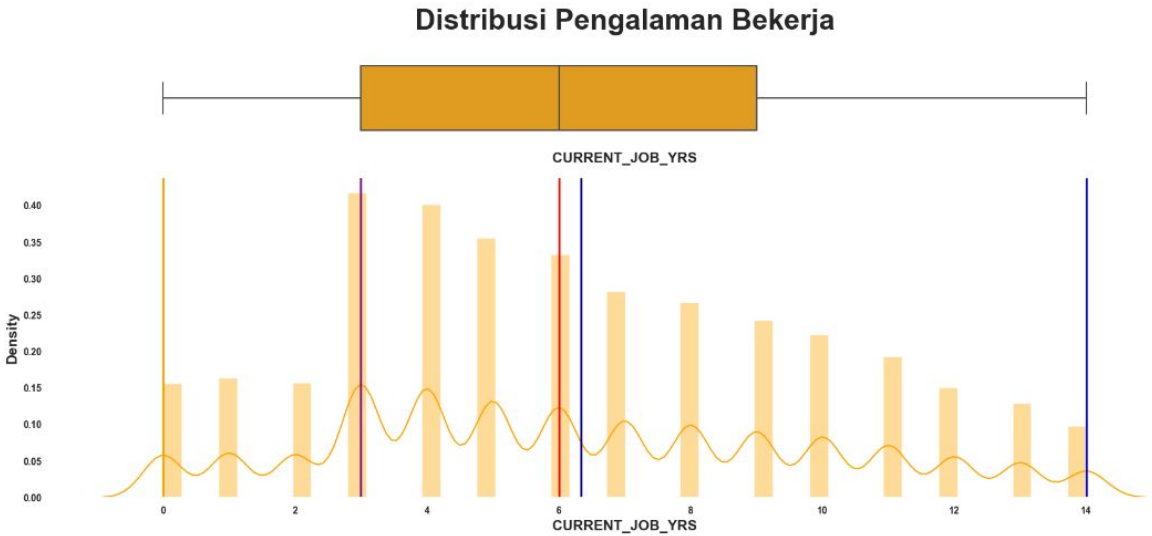
mean=10.1
median=10.0
max=20
min=0
mode=6



EXPERIENCE Column

UNIVARIATE ANALYSIS

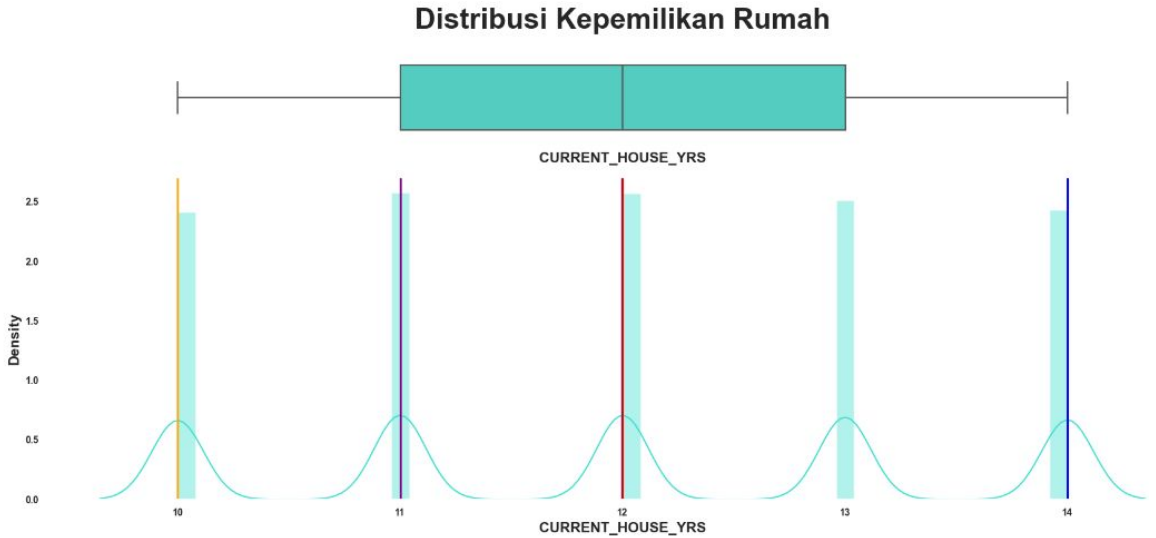
mean=6.3
median=6.0
max=14
min=0
mode=3



JOB YEARS Column

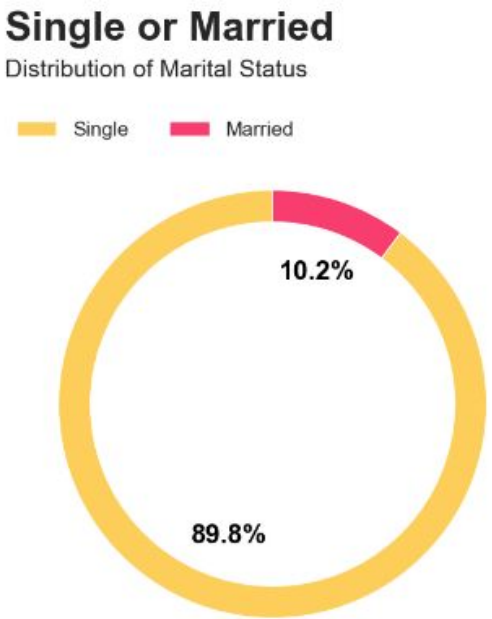
UNIVARIATE ANALYSIS

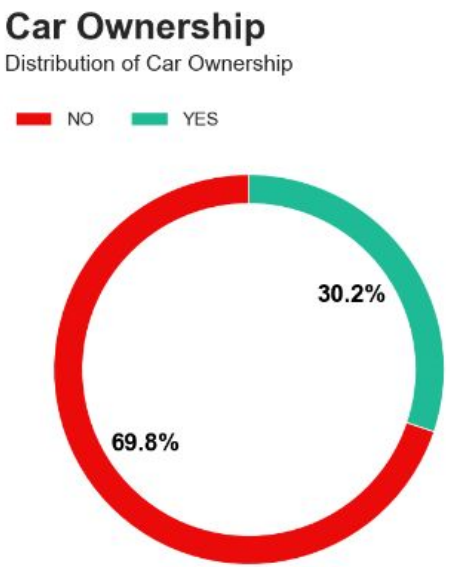
mean=12.0
median=12.0
max=14
min=10
mode=11

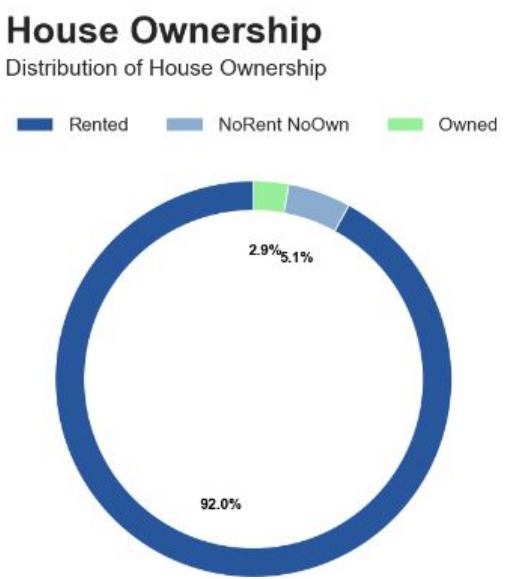


HOUSE YEARS Column

UNIVARIATE ANALYSIS







What can we get from univariate analysis?

AGE

- Rata-rata umur pelanggan adalah 50 tahun, dengan median dan mode yang sama pada 50 tahun.
- Umur pelanggan bervariasi antara 21 hingga 79 tahun.

INCOME

- Rata-rata pendapatan pelanggan adalah sekitar 49.971.161, sedangkan mediannya adalah sekitar 5.000.694,5.
- Pendapatan pelanggan berkisar dari 10.310 hingga 9.999.938.

EXPERIENCE

- Rata-rata pengalaman profesional pelanggan adalah sekitar 10,1 tahun, dengan median dan mode pada 10 tahun.
- Pengalaman profesional berkisar dari 0 hingga 20 tahun.

CURRENT_JOB_YRS

- Rata-rata tahun bekerja saat ini pelanggan adalah sekitar 6,3 tahun, dengan median pada 6 tahun dan mode pada 3 tahun.
- Tahun bekerja saat ini berkisar dari 0 hingga 14 tahun.

CURRENT_HOUSE

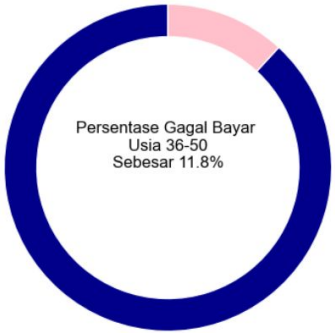
- Rata-rata tahun tinggal saat ini pelanggan adalah sekitar 12 tahun, dengan median dan mode pada 12 tahun.
- Tahun tinggal saat ini berkisar dari 10 hingga 14 tahun.

CUST_CAT DATA

- Sebagian besar pelanggan (89.79%) adalah single.
- Mayoritas pelanggan (sekitar 92.02%) menyewa rumah.
- Mayoritas pelanggan (sekitar 69.84%) tidak memiliki mobil.

Bivariate Analysis

21-35 Tahun



36-50 Tahun

51-65 Tahun

66-79 Tahun

Based On Age Group

Interpretations:

Kelompok rentang usia 21-35 tahun (usia produktif) memiliki kecenderungan 2% lebih besar gagal bayar dibanding kelompok usia lainnya.

Kelompok rentang usia 66-79 tahun (usia lanjut) memiliki kecenderungan risiko gagal bayar yang paling kecil.

Semakin besar rentang usia nya maka semakin menurun risiko gagal bayarnya.

Conclusion:

Seiring bertambahnya usia, seseorang mungkin memiliki lebih banyak waktu untuk membangun stabilitas keuangan.

Bivariate Analysis

0-3 Tahun

3-6 Tahun

6-10Tahun

10-20 Tahun

Based On Experience Category

Interpretations:

Persentase gagal bayar kategori pengalaman 0-3 tahun 3% lebih besar dari kategori pengalaman 3-6 tahun dan 4% lebih besar dari kategori 5-10 dan 10-20 tahun.

Semakin besar kategori pengalaman-nya nya maka semakin kecil risiko gagal bayarnya.

Conclusion:

Seseorang dengan pengalaman kerja yang lebih lama mungkin memiliki stabilitas keuangan yang lebih tinggi.

Bivariate Analysis

Based On Income Category

Very Low Income

Very High Income

Moderate Income

High Income

Low Income

Interpretations:

Persentase gagal bayar terbesar ada pada Kategori Very Low Income (13.2%) dan disusul Kategori Very High Income (12.7%).

Ternyata customer dengan Kategori Low Income memiliki risiko kredit paling kecil.

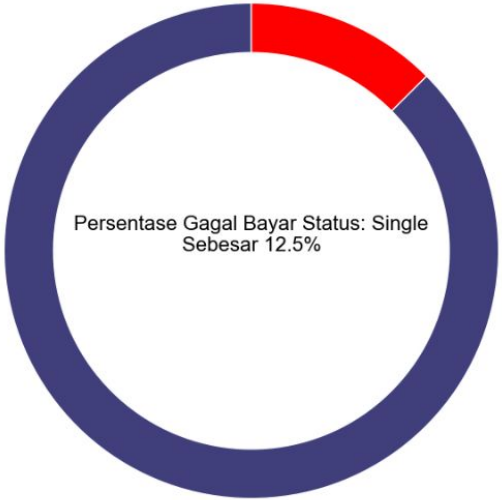
Tidak ada korelasi yang jelas antara tingkat pendapatan dan kemungkinan memiliki risiko dan semua kategori pendapatan memiliki persentase yang relatif serupa.

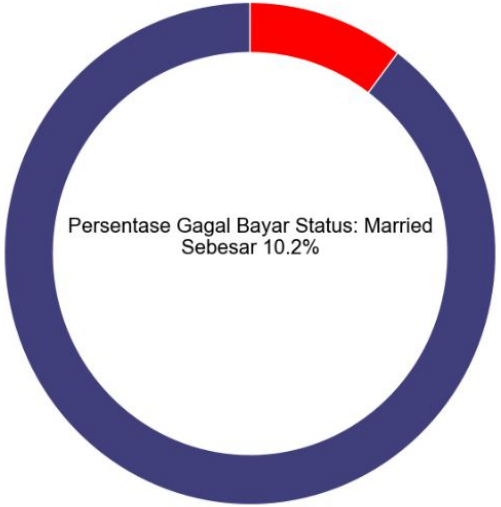
Bivariate Analysis

Based On Marital Status

Single

Persentase risiko gagal bayar pada individu yang status perkawinannya single sekitar 2.5% lebih tinggi daripada pada individu yang status perkawinannya married.



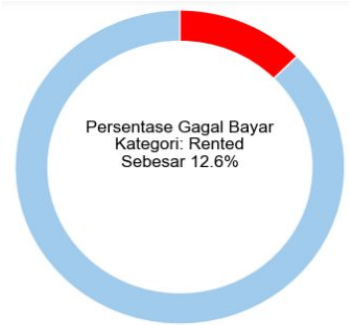


Married

Individu yang status perkawinannya Married cenderung memiliki risiko gagal bayar yang sedikit lebih rendah daripada individu yang status perkawinannya Single..

Bivariate Analysis

Based On House Ownership Status

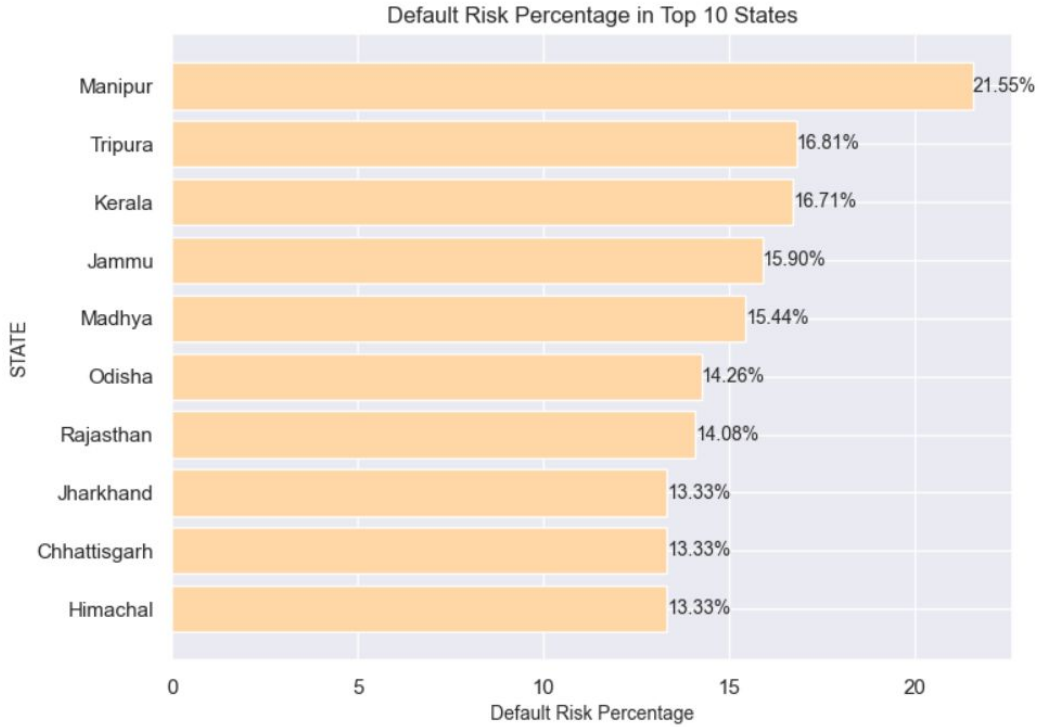


Individu yang tinggal di rumah sewa cenderung memiliki persentase risiko gagal bayar yang lebih tinggi (sekitar 2.6%) dibandingkan dengan individu yang tinggal di rumah milik dan tidak memiliki kepemilikan (sekitar 3.6%).

Kepemilikan rumah mungkin memiliki pengaruh pada risiko gagal bayar, dengan individu yang tinggal di rumah sewa cenderung memiliki risiko gagal bayar yang lebih tinggi daripada individu yang memiliki rumah.

Bivariate Analysis

Based On States



Interpretations:

Negara bagian Manipur menjadi state dengan risiko gagal bayar tertinggi sebesar 22%, diikuti Tripura dan Kerala sebesar 17%.

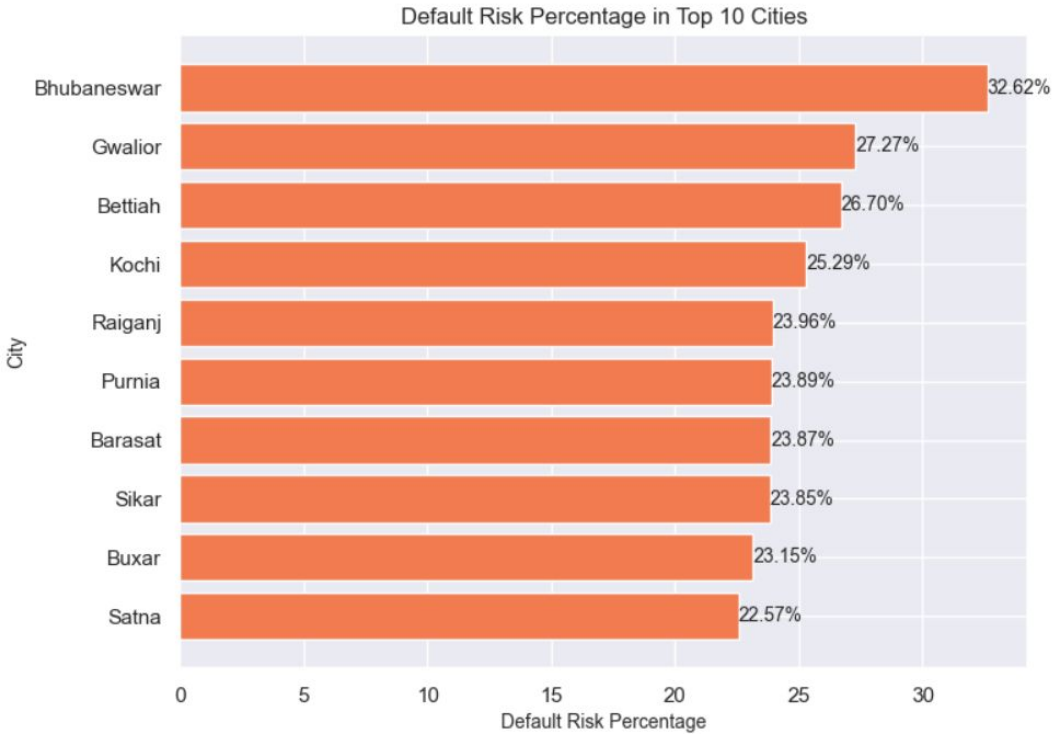
Rata-rata persentase risiko gagal bayar adalah 12.5%.

Hanya sekitar 10% wilayah negara bagian yang memiliki risiko dibawah 10% (Punjab, Uttarakhand, Chandigarh, Sikkim).

Hanya negara bagian Manipur yang memiliki risiko kredit lebih dari 20%.

Bivariate Analysis

Based On Cities



Interpretations:

Kota Bhubaneswar menjadi kota dengan risiko gagal bayar tertinggi sebesar 33%.

Rata-rata persentase risiko gagal bayar adalah 25.31%.

Kota Bhubaneswar, Gwalior, Bettiah, dan Kochi merupakan kota dengan seperempat lebih (>25%) penduduknya mengalami gagal bayar.

Nilai std pada pcg sebesar 2.971649 menunjukkan bahwa ada variasi yang cukup signifikan dalam persentase risiko gagal bayar di antara kota-kota.

Multivariate Analysis

Current_Job-Years and Experience

Terdapat korelasi positif yang cukup kuat antara ‘Experience' dengan 'CURRENT_JOB_YRS',

Hal ini menunjukan mungkin pengalaman kerja akan meningkat seiring bertambahnya lama bekerja.

Risk_Flag

Terdapat korelasi negatif antara 'Risk_Flag' dengan 'Experience', 'CURRENT_JOB_YRS', dan ‘Age’.

Hal ini mengindikasikan bahwa semakin besar umur, pengalaman,dan lama bekerja maka semakin rendah kemungkinan terjadinya risiko gagal bayar.


2. Data Wrangling

Data Wrangling

Procesess

01

Ekstraksi Kota dan Negara Bagian: Dilakukan ekstraksi nama kota dan negara bagian dari kolom CITY dan STATE.

02

- Feature_Engineering: Age_Group, Income_Category, Experience_Category, profession_map, dan zone.
- Encoding: Label Encoder Method.

03

- ANOVA Test.
- Chi Square Test.

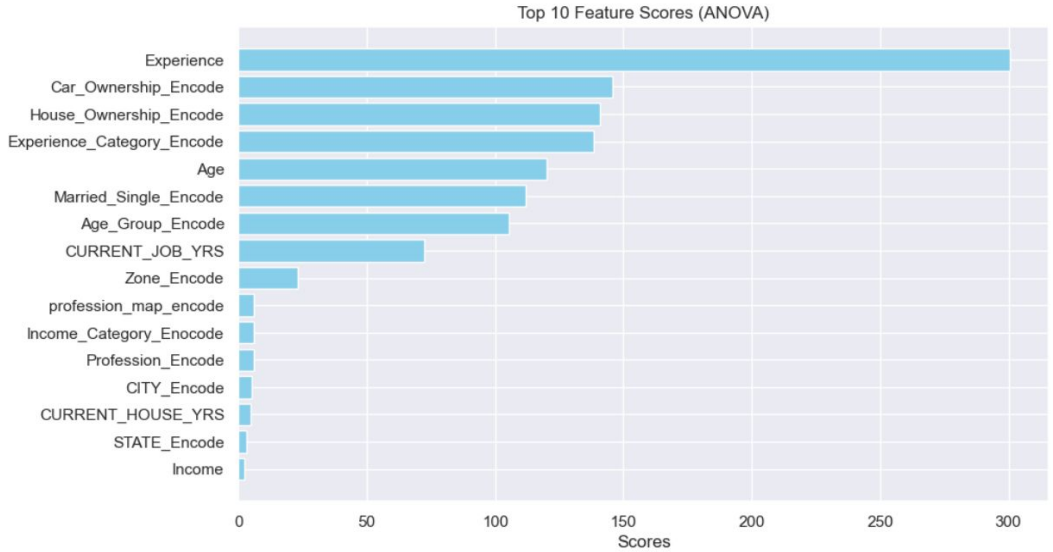
04

Handle Class Imbalance pada fitur target (Risk_Flag).

FEATURE SELECTION

Hasil Anova test memberikan pandangan untuk mengevaluasi fitur-fitur mana saja yang relevan terhadap variabel target. Dari hasil tersebut, beberapa fitur yang menonjol adalah Experience, Age, CURRENT_JOB_YRS, Income, dan House_Ownership_Encode

ANOVA TEST RESULT

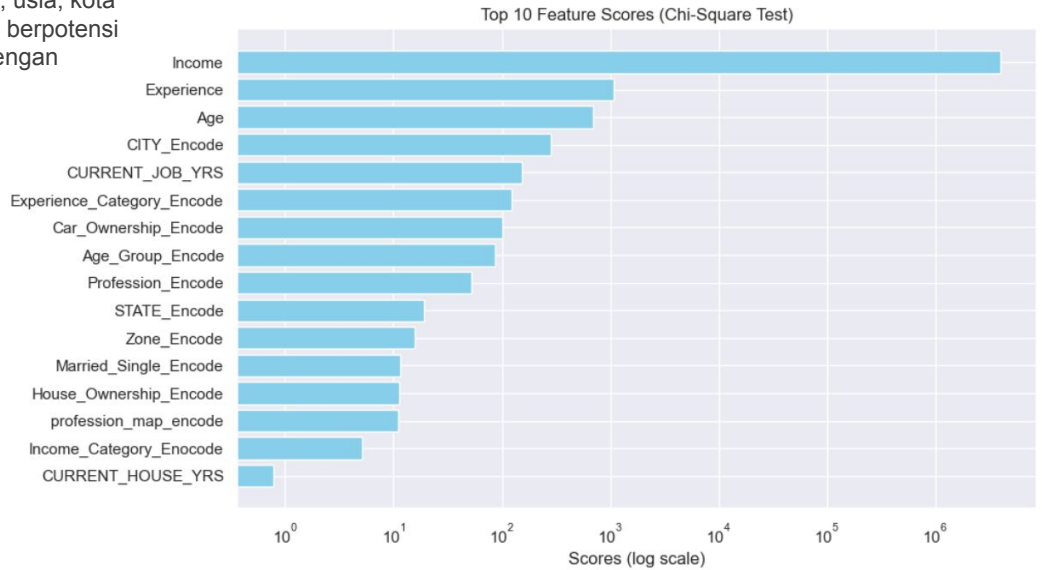


ANOVA Test

FEATURE SELECTION

Dari hasil test, terlihat bahwa fitur-fitur yang paling signifikan dalam kaitannya dengan Risk_Flag adalah Income, Experience, Age, CITY_Encode, dan CURRENT_JOB_YRS. Hal ini menunjukkan bahwa tingkat pendapatan, pengalaman, usia, kota tempat tinggal, dan lama bekerja saat ini berpotensi memiliki korelasi yang signifikan dengan kemungkinan gagal bayar.

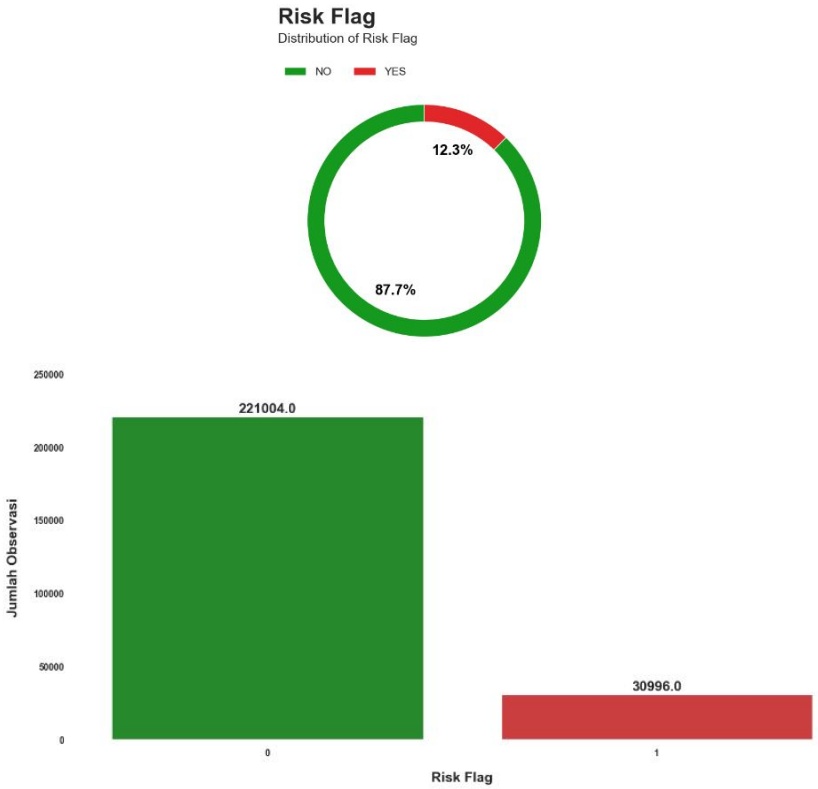
CHI-SQUARE TEST RESULT



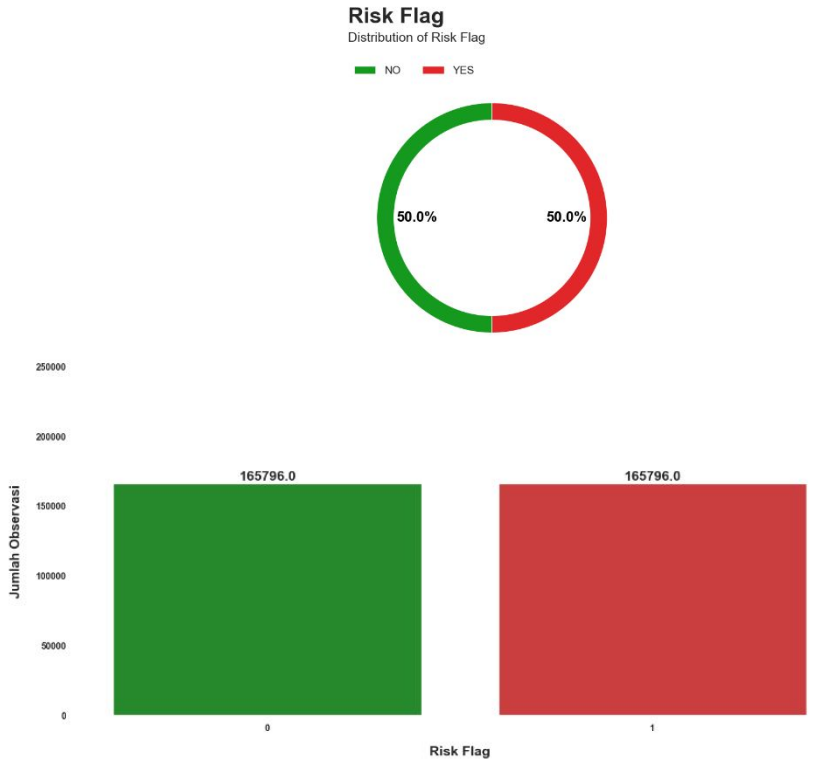
CHI-SQUARE Test

Handling Imbalance Data

Before



After



3. Modelling

Machine Learning Evaluation

Evaluation	Logistic
Train Accuracy	57.6%
Test Accuracy	55.4%
Avg F1-Score	46%
Avg CV-Score	57.5%

Evaluation	XGBoost
Train Accuracy	89.3%
Test Accuracy	87%
Avg F1-Score	75%
Avg CV-Score	

Evaluation	DecisionTree
Train Accuracy	93.8%
Test Accuracy	86.1%
Avg F1-Score	76%
Avg CV-Score	90.1%

Evaluation	RandomForest
Train Accuracy	95.8%
Test Accuracy	88.5%
Avg F1-Score	78%
Avg CV-Score	93.4%

Logistic Regression

Best Parameters:

{'C': 0.001, 'penalty': 'l1', 'solver': 'saga'}

XGBoost Classifier

Best Parameters:

{'n_estimators': 300, 'max_depth': 7, 'learning_rate': 0.1}

Decision Tree

Best Parameters:

{'splitter': 'best', 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': sqrt, 'max_depth': 50, 'criterion': 'entropy'}

Random Forest

Parameters:

Default Parameters

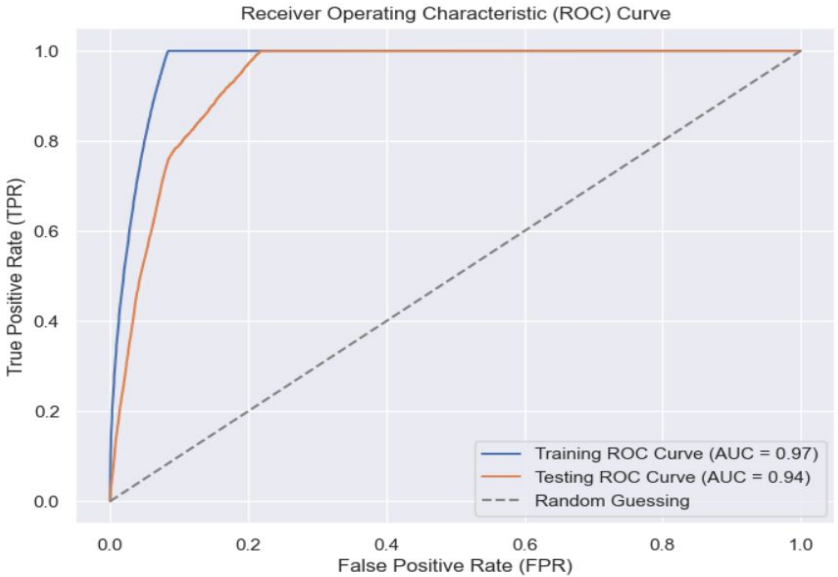
Random Forest Classifier

Why Random Forest?

Berdasarkan evaluasi model, secara umum, RandomForest memiliki performa yang baik berdasarkan evaluasi yang diberikan, dengan nilai Akurasi, F1-Score, dan CV-Score yang tinggi.

Avg_Precision: 75%

Avg_Recall: 85%



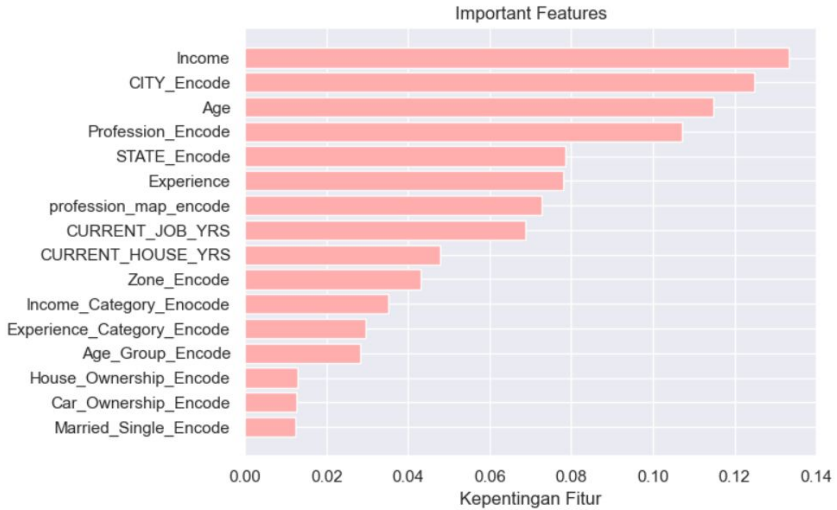
ROC Curve Results

Dari hasil yang diberikan, terlihat bahwa model memiliki kinerja yang sangat baik baik pada data pelatihan maupun data pengujian. Area di bawah kurva untuk data pelatihan adalah 0.97, sementara untuk data pengujian adalah 0.94. Ini menunjukkan bahwa model memiliki kemampuan yang baik untuk memisahkan kelas positif dan negatif, serta memiliki kemampuan yang baik dalam generalisasi ke data yang tidak terlihat sebelumnya (data pengujian). Oleh karena itu, dapat disimpulkan bahwa model memiliki kinerja yang baik dan dapat diandalkan untuk memprediksi kelas target.

Feature Importance

Based On Decision Tree and Random Forest

Berdasarkan hasil analisis feature importance, kita dapat mengevaluasi kontribusi relatif dari setiap fitur terhadap kinerja model. Fitur-fitur yang paling signifikan adalah Pendapatan (Income), Kota (CITY_Encode), Usia (Age), Profesi (Profession_Encode), dan Negara Bagian (STATE_Encode). Fitur-fitur ini memiliki dampak yang paling besar dalam memprediksi target dalam model.



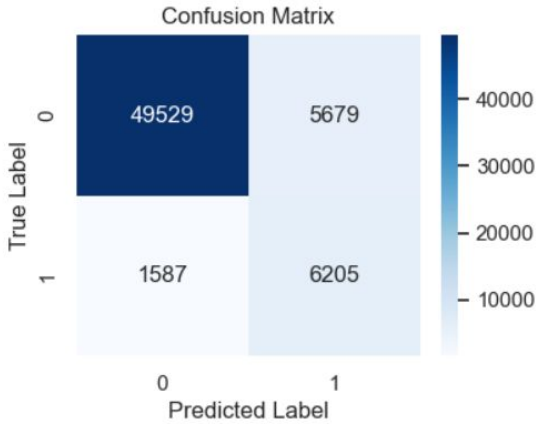
Random Forest

Assumption Calculation

If assumed there are 63.000 borrowers with the same loan amount all 1 Lakh (100,000 Rupees).

Calculation Steps:

- Peminjam yang diberikan pinjaman berdasarkan model (FN + TN): 1.587 + 49.529 = 51.116 peminjam.
- Total peminjam yang diberikan: 51.116 peminjam.
- Total Peminjam gagal bayar (Default): 1.587 (FN).
- Persentase Repayment Rate = (Pengembalian / Total Peminjam) * 100% = (49.529 Peminjam / 51.116 Peminjam) * 100% ≈ 96.9%.



Estimated success of loan 'Repayment Rate' using Machine Learning Model

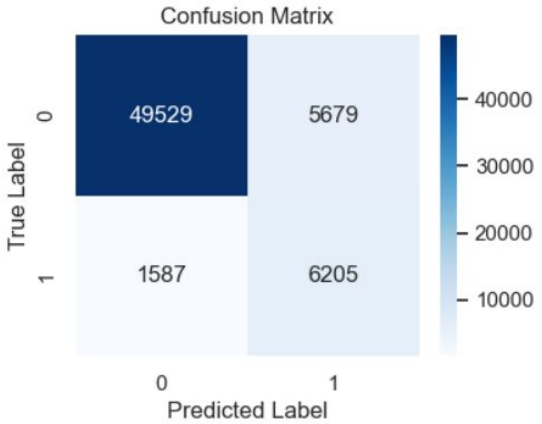
96.9%

Assumption Calculation

Comparison of Manual Accuracy and Using Machine Learning

Calculation Steps:

- Peminjam yang diberikan pinjaman = 63.000 peminjam.
- Total Peminjam gagal bayar (Default): 7.793 (FN + TP).
- Persentase Repayment Rate = (Pengembalian / Total Peminjam) * 100% = (55.207 Peminjam / 63.000 Peminjam) * 100% ≈ 87.6%.

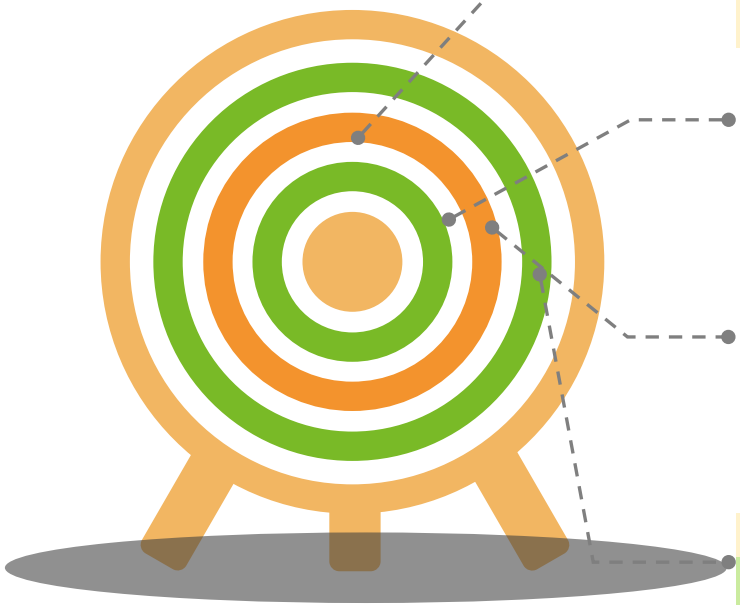


The loan 'Repayment Rate' increased significantly:

87.6% 96.9%

4. Business Recommendation

Business Recommendation



Segmentasi Usia dan Penyesuaian Strategi

Fokus pada kelompok usia produktif (21-35 tahun) untuk memberikan perhatian khusus dalam pemantauan dan manajemen risiko, sementara memberikan penawaran khusus kepada kelompok usia lanjut (66-79 tahun) yang memiliki risiko gagal bayar yang lebih rendah.

Penilaian Status Perkawinan dan Kepemilikan Rumah

Bisnis dapat memberikan perhatian khusus pada individu yang status perkawinannya single dan mereka yang tinggal di rumah sewa, karena cenderung memiliki risiko gagal bayar yang lebih tinggi. Ini dapat mencakup penyesuaian tingkat bunga atau persyaratan lainnya untuk mengurangi risiko kredit.

Evaluasi Regional

Bisnis harus mempertimbangkan faktor regional dalam menilai risiko kredit. Negara bagian seperti Manipur, Tripura, dan Kerala memiliki risiko gagal bayar yang tinggi, sementara kota seperti Bhubaneswar memiliki risiko yang lebih tinggi dibandingkan dengan wilayah lainnya. Strategi pemasaran dan pengelolaan risiko harus disesuaikan dengan kondisi regional yang spesifik.

Optimasi Fitur dan Pengembangan Model

Fokus pada fitur-fitur yang paling signifikan seperti pendapatan, usia, dan lokasi dapat membantu meningkatkan akurasi prediksi dan pengelolaan risiko secara keseluruhan.

Conclusions

Analisis data pelanggan menunjukkan bahwa faktor-faktor seperti usia, pendapatan, pengalaman, status perkawinan, state, kota, dan kepemilikan rumah berpengaruh signifikan terhadap risiko kredit.

Rekomendasi bisnis meliputi segmentasi pelanggan yang lebih tepat dan penerapan strategi pemberian pinjaman yang lebih cerdas berdasarkan prediksi risiko kredit menggunakan model Machine Learning, sehingga dapat mengurangi risiko pinjaman dan meningkatkan tingkat pembayaran kembali.

Thank you for your Attention

Members of the group:

1. Dera Yuliani
2. Ichsan Abdilah Bimantara Pulungan
3. M. Fathul Radhiansyah
4. Muhammad Fatoni
5. Muhammad Salman Alfarizi
6. Nadiar Almahira Permatasari
7. Nur Islamiati Sanusi

By: Pearson Explorer