Yohanes Wiliam Hadiprojo

00000063762

The answer to each question below, including the subject, object and data, MUST BE THE SAME as the industry field and company clients you chose when answering the last Midterm Exam/ UTS question.

*For the students who at the past UTS did not choose the company's client, please make your choice according to the following business fields:*

A. Sport Factory Outlet

B. Motorbike Dealer (sales)

C. Agricultural business

D. Pharmaceutical Company

E. Banking

State the field of business you choose and give the company name “Bank/ Blackrock.” to answer the following questions.

**LINK DATASET:**

[**https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers**](https://www.kaggle.com/datasets/mathchi/churn-for-bank-customers)

**LINK SAS:** [**https://v4e062.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2F6ab61e14-a93a-45cd-b557-ec2b98b101ef&page=vi103**](https://v4e062.vfe.sas.com/links/resources/report?uri=%2Freports%2Freports%2F6ab61e14-a93a-45cd-b557-ec2b98b101ef&page=vi103)

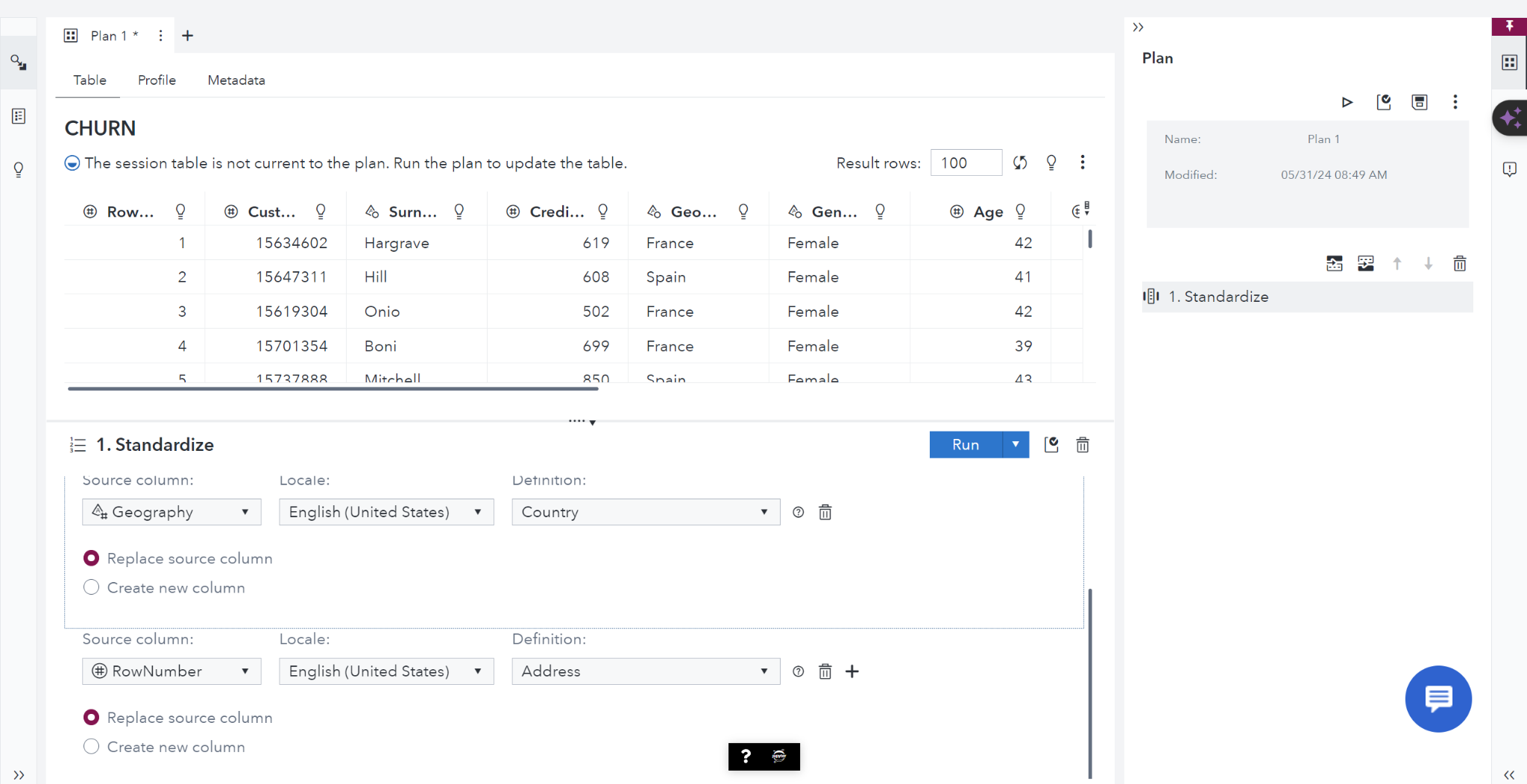
**Data overview**

Dataset yang digunakan dalam penelitian ini terdiri dari catatan nasabah bank dan atribut terkait, yang diperoleh dari [berikan sumber data atau jelaskan bagaimana dataset dikumpulkan]. Setiap catatan dalam dataset mewakili contoh nasabah yang unik, dengan fitur-fitur yang mencakup informasi demografis, perilaku perbankan, dan atribut rekening. Dataset berisi kolom-kolom berikut:

1. Row Number: Sesuai dengan nomor catatan (baris) dan tidak berpengaruh pada output.
2. CustomerId: Berisi nilai acak dan tidak berpengaruh pada nasabah yang meninggalkan bank.
3. Surname: Nama belakang nasabah, yang tidak berpengaruh pada keputusan mereka untuk meninggalkan bank.
4. CreditScore: Merupakan skor kredit nasabah, yang berpotensi mempengaruhi perpindahan nasabah, karena nasabah dengan skor kredit yang lebih tinggi memiliki kemungkinan yang lebih kecil untuk meninggalkan bank.
5. Geography: Menunjukkan lokasi nasabah, yang dapat mempengaruhi keputusan mereka untuk meninggalkan bank.
6. Gender: Menunjukkan jenis kelamin nasabah, yang menarik untuk dieksplorasi terkait perannya dalam perpindahan nasabah.
7. Age: Menunjukkan usia nasabah, yang relevan karena nasabah yang lebih tua umumnya lebih kecil kemungkinannya untuk meninggalkan bank mereka dibandingkan nasabah yang lebih muda.
8. Tenure: Mengacu pada jumlah tahun nasabah telah menjadi nasabah bank, dengan nasabah yang lebih tua biasanya lebih loyal dan kecil kemungkinannya untuk meninggalkan bank.
9. Balance: Mewakili saldo di rekening nasabah, yang berfungsi sebagai indikator signifikan dari perputaran nasabah, karena nasabah dengan saldo yang lebih tinggi cenderung tidak akan keluar dibandingkan dengan nasabah yang memiliki saldo yang lebih rendah.
10. NumOfProduct: Menunjukkaan jumlah produk yang telah dibeli nasabah melalui bank.
11. HasCrCard: Menunjukkan apakah nasabah memiliki kartu kredit atau tidak, yang relevan karena nasabah yang memiliki kartu kredit lebih kecil kemungkinannya untuk meninggalkan bank.
12. IsActiveMember: Menunjukkan apakah nasabah merupakan anggota aktif, karena nasabah yang aktif cenderung tidak akan meninggalkan bank.
13. Estimated Salary: Mewakili perkiraan gaji nasabah, yang berfungsi sebagai indikator lain dari perputaran nasabah, karena nasabah dengan gaji yang lebih rendah lebih mungkin untuk keluar dibandingkan dengan mereka yang memiliki gaji yang lebih tinggi.
14. Exited: Variabel biner yang menunjukkan apakah nasabah keluar dari bank atau tidak.

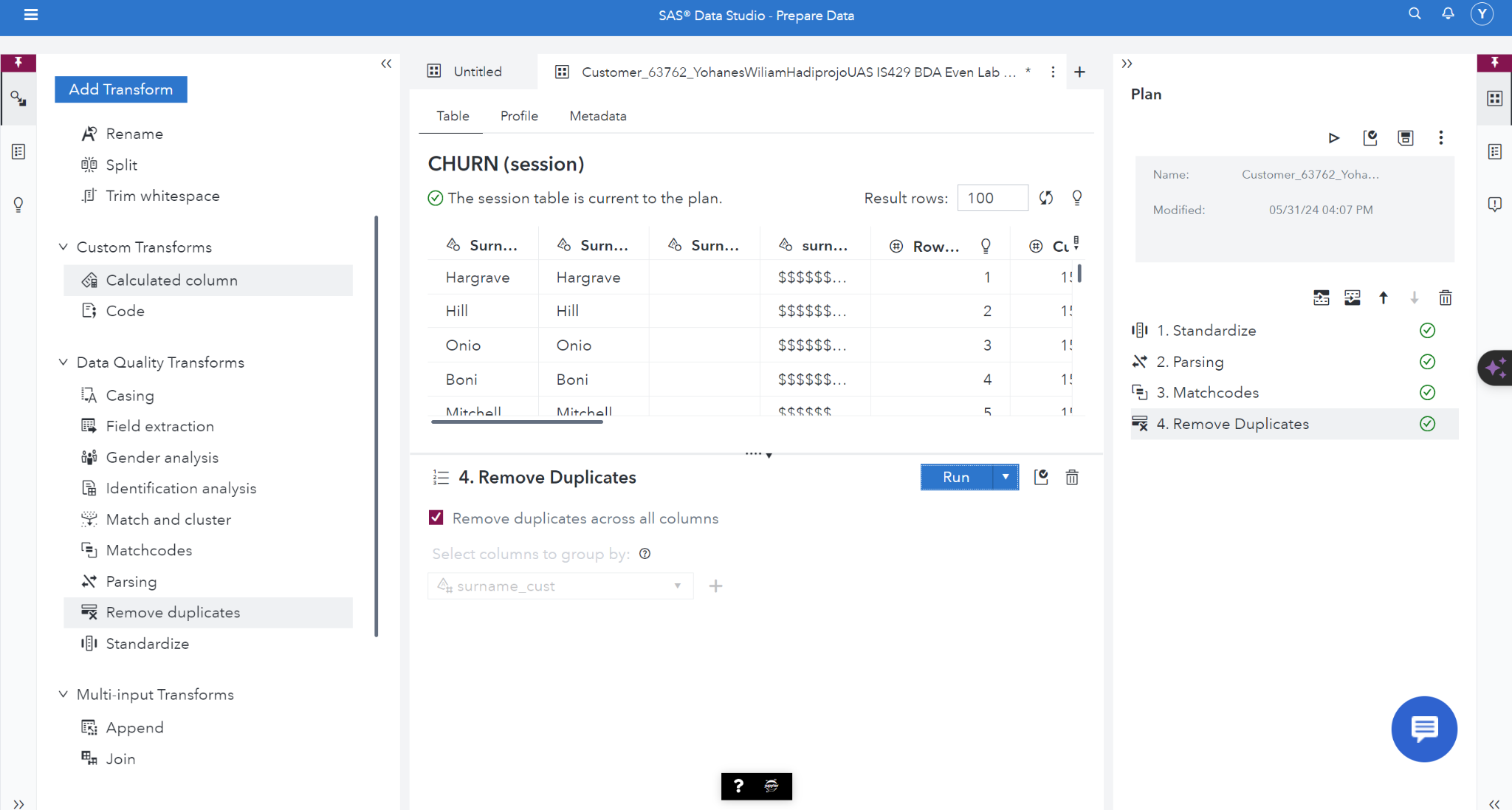
**Data Preparation**

**standarize**



Menstandarisasi Variabel Numerik: Variabel numerik seperti Exited, Geography, Row Number distandarisasi agar memiliki nilai rata-rata 0 dan deviasi standar 1 untuk memastikan penskalaan yang seragam di seluruh fitur.

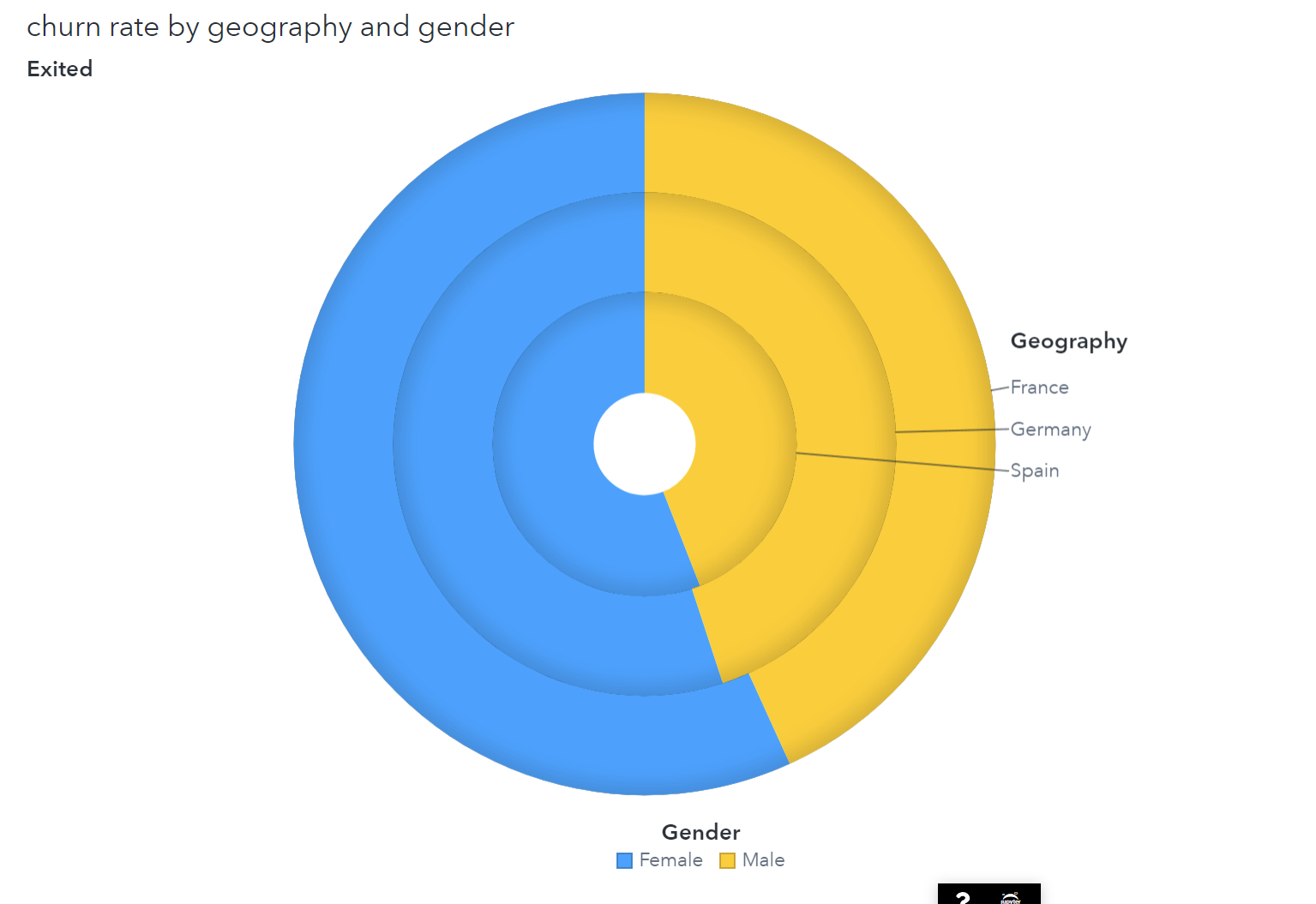
**Removes duplicates**



Handling missing value

Setiap nilai yang hilang dalam dataset diperhitungkan menggunakan teknik yang sesuai seperti imputasi rata-rata atau imputasi prediktif.

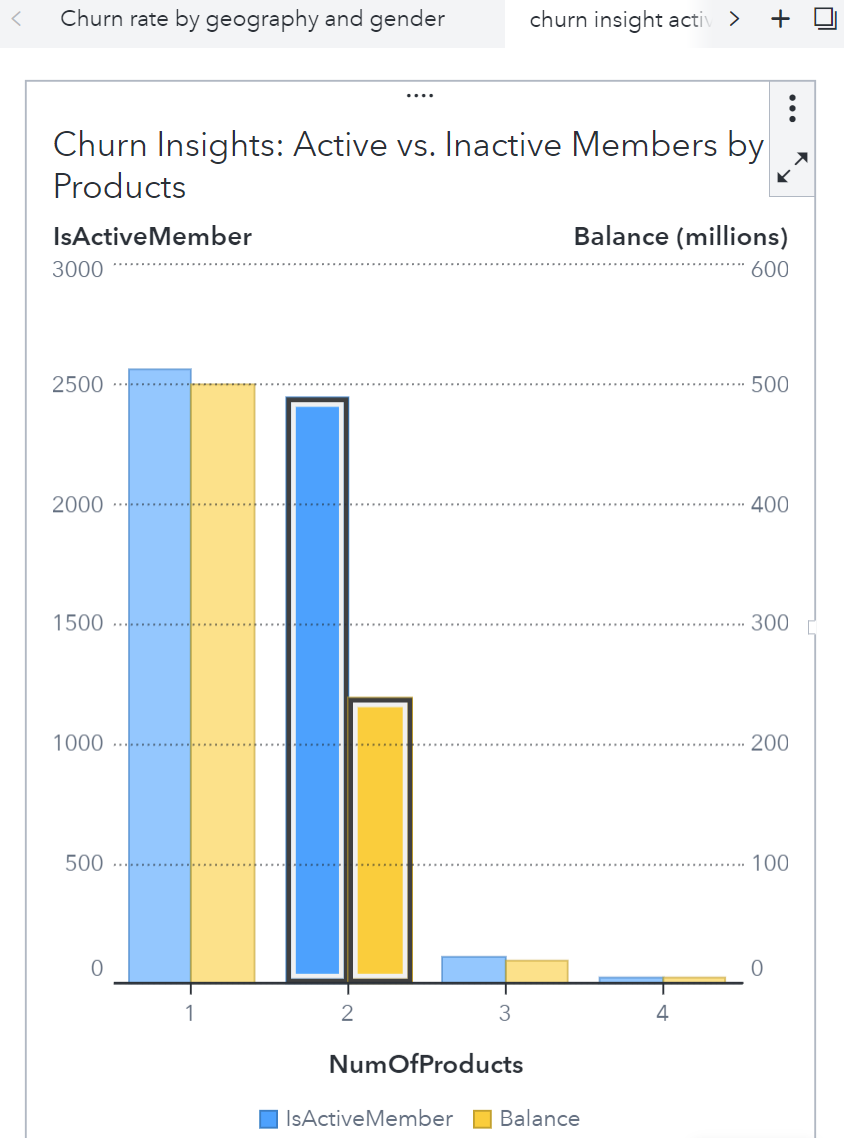
**EDA**



Fase analisis data eksplorasi melibatkan visualisasi dan analisis tingkat churn berdasarkan geografi dan gender untuk mengungkap pola dan wawasan. Kumpulan data mencakup tiga wilayah geografis: Prancis, Jerman, dan Spanyol. Tingkat churn divisualisasikan menggunakan bagan donat, yang mengungkapkan beberapa wawasan utama. Untuk setiap geografi, segmen bagan dibedakan dengan dua warna: biru mewakili pelanggan pria dan kuning mewakili pelanggan wanita. Lingkaran luar menggambarkan tingkat churn total untuk setiap geografi, sedangkan lingkaran dalam membedakan tingkat churn berdasarkan jenis kelamin.

Analisis ini menyoroti perbedaan yang mencolok dalam tingkat churn di seluruh wilayah geografis dan gender. Di Prancis, tingkat churn rate perempuan adalah 55,0%, dengan 448 orang yang keluar, dibandingkan dengan tingkat churn rate laki-laki sebesar 44,1%, dengan 366 orang yang keluar. Jerman menunjukkan tren yang sama, dengan tingkat churn rate perempuan sebesar 59,9% (231 orang keluar) dan tingkat churn rate laki-laki sebesar 43,2% (182 orang keluar). Di Spanyol, tingkat keluar masuknya karyawan perempuan mencapai 55,5%, sementara tingkat keluar masuknya karyawan laki-laki mencapai 43,7%, dengan 350 karyawan laki-laki yang keluar. Namun, jumlah pasti individu perempuan yang keluar di Spanyol tidak dapat ditranskrip karena adanya potensi kesalahan.

Temuan ini menunjukkan bahwa gender memainkan peran penting dalam perpindahan pelanggan, dengan pelanggan wanita secara konsisten menunjukkan tingkat perpindahan yang lebih tinggi daripada pelanggan pria di seluruh wilayah. Selain itu, perbedaan regional juga terlihat jelas, dengan Jerman menunjukkan tingkat churn tertinggi secara keseluruhan. Wawasan ini menunjukkan perlunya bank mengembangkan strategi retensi yang ditargetkan yang memperhitungkan variasi gender dan geografis untuk mengurangi churn secara efektif. EDA awal ini memberikan gambaran yang berharga, dan analisis lebih lanjut dapat mempelajari faktor-faktor tambahan yang mempengaruhi churn untuk menginformasikan strategi retensi yang lebih bernuansa.



Dalam analisis data eksplorasi kami, kami memeriksa tingkat churn berdasarkan geografi dan jenis kelamin dengan menggunakan visualisasi diagram batang yang berjudul "Tingkat Churn Berdasarkan Geografi dan Jenis Kelamin Menggunakan Anggota Aktif yang Tidak Aktif Berdasarkan Produk." Visualisasi ini memberikan wawasan yang berharga tentang hubungan antara churn pelanggan, penggunaan produk, dan status aktivitas anggota.

Sumbu x pada diagram batang menunjukkan jumlah produk yang digunakan oleh nasabah bank, mulai dari 1 (menunjukkan nasabah yang hanya menggunakan satu produk) hingga 4+ (menunjukkan nasabah yang menggunakan empat produk atau lebih). Sumbu y di sebelah kiri menunjukkan jumlah anggota, yang membedakan antara anggota yang aktif dan tidak aktif, sedangkan sumbu y di sebelah kanan menunjukkan saldo (dalam jutaan) yang terkait dengan setiap jumlah produk.

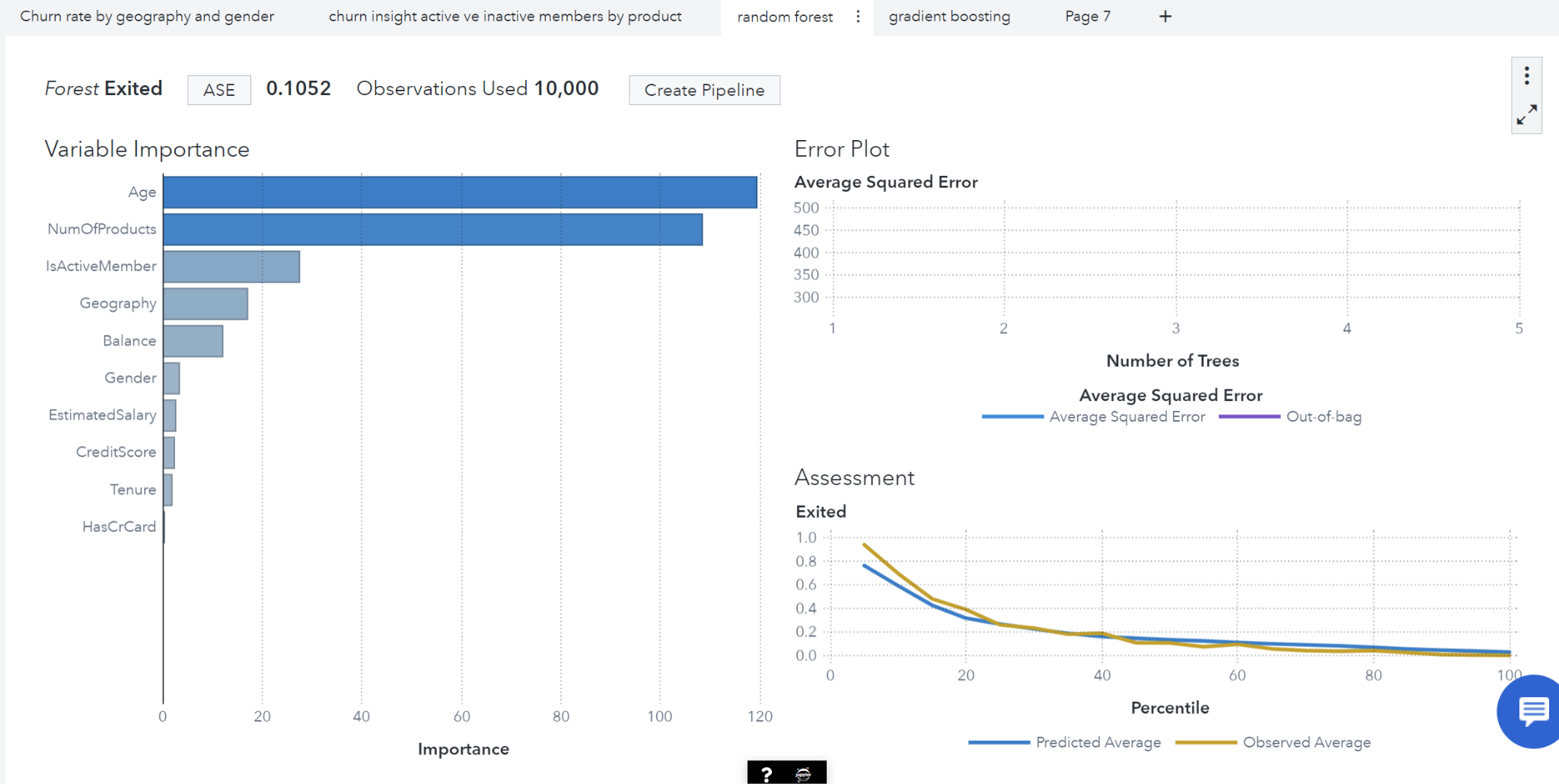
Beberapa wawasan penting dapat diambil dari visualisasi ini. Pertama, nasabah dengan lebih banyak produk cenderung memiliki saldo yang lebih tinggi, yang mengindikasikan korelasi positif antara jumlah produk yang digunakan nasabah dengan saldo keseluruhan. Kedua, tingkat churn rate lebih tinggi pada nasabah yang memiliki lebih sedikit produk, terutama nasabah yang hanya memiliki 1 atau 2 produk. Hal ini menunjukkan bahwa nasabah yang menggunakan lebih sedikit produk bank lebih cenderung meninggalkan bank.

Untuk mengurangi churn, bank harus fokus untuk mempertahankan nasabah yang memiliki banyak produk, karena nasabah ini tidak hanya memiliki saldo yang lebih tinggi tetapi juga lebih kecil kemungkinannya untuk berpindah bank. Pendekatan yang ditargetkan ini dapat membantu bank meningkatkan loyalitas nasabah dan mengurangi tingkat churn secara keseluruhan.

Meskipun visualisasi ini memberikan gambaran hubungan antara penggunaan produk dan churn, analisis lebih lanjut dapat mengeksplorasi faktor-faktor lain yang mempengaruhi churn untuk mengembangkan strategi retensi yang lebih komprehensif.

**Modeling**

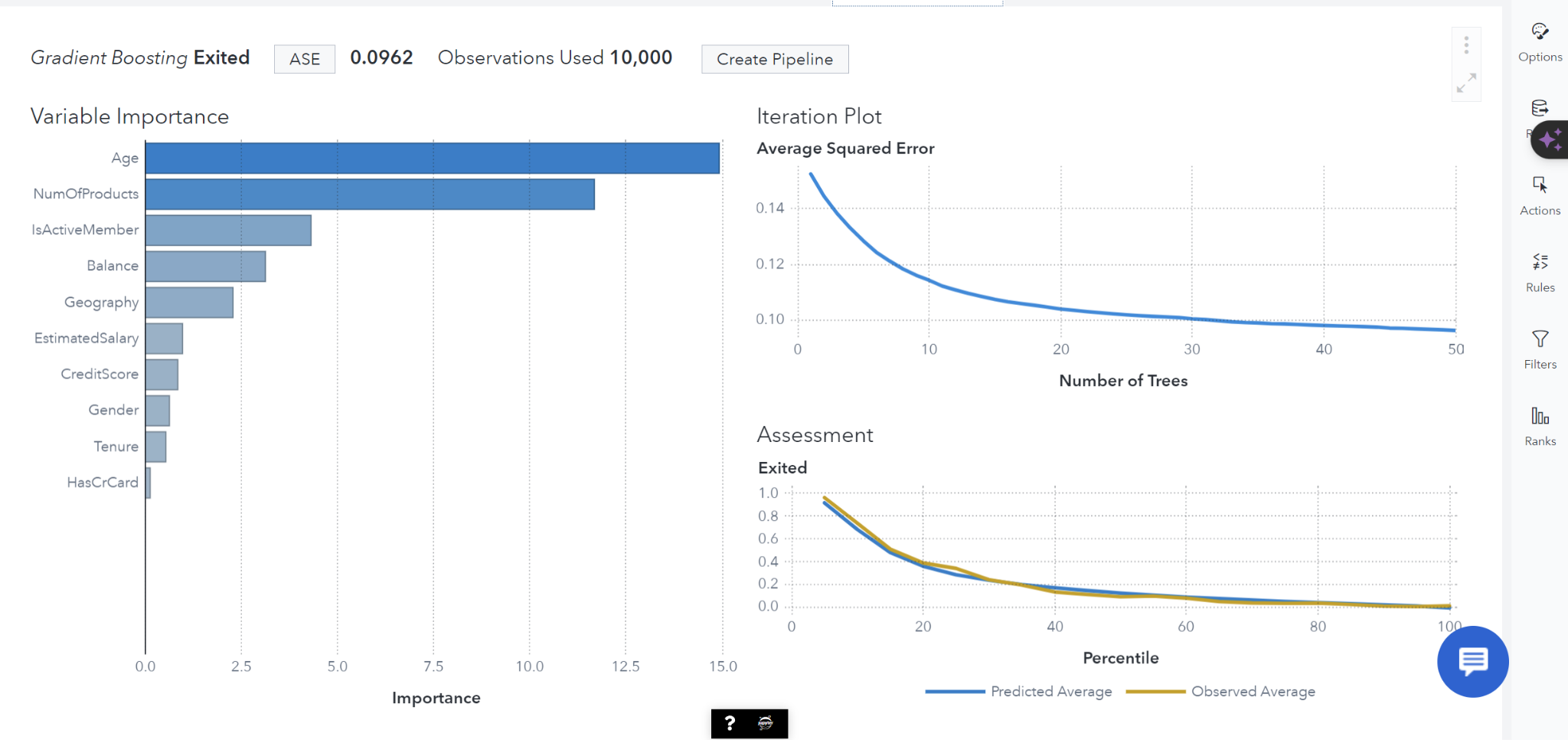
**Random forest**



Hasil dari model Random Forest memperkuat beberapa wawasan dari analisis data eksplorasi sebelumnya:

* Keterlibatan Pelanggan: Fitur-fitur seperti jumlah produk yang digunakan dan status keanggotaan aktif sangat penting untuk memprediksi churn, yang menekankan pentingnya keterlibatan pelanggan dalam strategi retensi.
* Faktor Demografis: Usia dan geografi merupakan prediktor yang signifikan, menunjukkan bahwa strategi retensi harus disesuaikan dengan kelompok usia dan wilayah yang berbeda.
* Faktor Keuangan: Meskipun saldo dan perkiraan gaji penting, namun tingkat kepentingannya yang lebih rendah dibandingkan dengan fitur-fitur keterlibatan menunjukkan bahwa status keuangan saja bukanlah prediktor terkuat dari churn.

**Gradiend Boosting**



Hasil dari model Gradient Boosting memperkuat beberapa wawasan dari analisis data eksplorasi sebelumnya dan model Random Forest:

* Keterlibatan Pelanggan: Fitur-fitur seperti jumlah produk yang digunakan dan status keanggotaan aktif sangat penting untuk memprediksi churn, menyoroti pentingnya keterlibatan pelanggan dalam strategi retensi.
* Faktor Demografis: Usia dan geografi adalah prediktor yang signifikan, yang menunjukkan perlunya strategi retensi yang disesuaikan berdasarkan segmen demografis.
* Faktor Keuangan: Saldo dan perkiraan gaji merupakan hal yang penting, mendukung gagasan bahwa nasabah yang stabil secara finansial cenderung tidak berpindah.

Dengan memanfaatkan wawasan ini, bank dapat mengembangkan strategi yang lebih efektif untuk memitigasi churn, dengan fokus pada peningkatan keterlibatan nasabah, menyediakan layanan yang dipersonalisasi berdasarkan faktor demografis, dan mempertimbangkan perbedaan regional.