📖 Absa Customer Income Prediction – Conceptual Story

1️⃣ The Problem

- **Goal:** Predict the **net income** of each customer using **demographic and transaction data**.
- **Why it matters:** Helps the bank understand customers better, e.g., for credit scoring, product recommendations, or financial planning.

2️⃣ The Data

- 14 months of **transaction history** → contains thousands of individual transactions per customer.
- **Demographics** → age, employment, location, etc.
- **Target variable:** declared net income.
- **Challenge:** Data is high-dimensional, mixed-type (categorical + numerical), and has variable transaction frequencies per customer.

3️⃣ Why Random Forest Regression

- **Predicts continuous outcomes** (net income).
- **Captures nonlinear relationships:** Income is affected by complex patterns in transactions, not just a straight line.
- **Handles noisy and heterogeneous data**: Some customers have 1 transaction, others 2,000 — RF is robust.
- **Provides feature importance:** Bank can see which behaviors (transactions, demographics) matter most for income prediction.
- RF is strong and interpretable.

4️⃣ Feature Engineering – Turning Transactions into Signals

- **Problem:** Random Forest can't directly read thousands of raw transactions per customer.
- **Solution:** Aggregate the transactions into **summary features per customer**, e.g.:
    - Total transaction amount
    - Average transaction amount
    - Number of transactions
    - Max, min, std of transactions
    - Transaction types (deposits, withdrawals, payments)
- **Combine** these with demographics (age, employment status) → final dataset for modeling.

5️⃣ Key Random Forest Hyperparameters – Story Behind Choices

1. **n_estimators (number of trees)**
    a. More trees → more stable predictions. Reduces randomness.
2. **max_depth (maximum levels per tree) -Controls how deep each tree grows to avoid overfitting rare data."**
3.
    a. Controls complexity:
        i. Low depth → model is too simple → may **underfit** (miss important patterns).
        ii. High depth → model is too complex → may **overfit** (memorize rare customer behavior).
    b. Balances generalization vs memorization.
4. **min_samples_leaf (minimum samples per leaf)**
    a. Ensures leaves have enough data → prevents the tree from reacting to rare, extreme transactions.
    b. Works with max_depth to avoid overfitting.

6 Training & Validation

- Split train set into **train + validation** → ensures model generalizes to unseen customers.
- Feature importance can be extracted → helps explain **what drives income predictions**.

7 Evaluation

- Predict on test set → submit to competition.
- Performance metric usually measures **how close predicted income is to actual income** (e.g., RMSE, MAE).

8 The Big Picture

"We transform raw customer data into meaningful patterns, then use Random Forest Regression because it is robust, interpretable, and can capture complex relationships in income behavior, balancing accuracy and generalization."

"Our project predicts a customer's income using their **transaction history** and **demographic information**.

First, we turn the raw transaction data into **useful numbers**, like total spent, average transaction, and number of transactions, and combine that with things like age and employment status.

We use **Random Forest Regression** because it can handle messy data, learn **complex patterns**, and tell us which features are most important.

We control the model with settings like max_depth — this stops the trees from getting too complicated and overfitting, while still learning the important trends.

In short, we take messy customer data, summarize it into meaningful features, and use a model that's strong, reliable, and interpretable to predict income."

1️⃣ Basics / Concepts

Q: What is machine learning?
A: "It's a way for computers to learn patterns from data and make predictions without being explicitly programmed."

Q: What's the difference between supervised and unsupervised learning?
A:

- Supervised → "We know the answers (labels) and train the model to predict them."
- Unsupervised → "We don't have labels; the model finds patterns or groups in the data."

Q: What is overfitting?
A: "When a model learns the training data too well, including noise, so it performs poorly on new data."

Q: What is underfitting?
A: "When a model is too simple and cannot capture the important patterns in the data."

2️⃣ Models / Algorithms

Q: Why use Random Forest instead of Linear Regression?
A: "Random Forest can capture complex, nonlinear relationships and handle noisy or mixed-type data, whereas linear regression assumes a straight-line relationship."

Q: What's the difference between Random Forest and Decision Trees?
A: "Random Forest is a group of decision trees that vote together. This makes it more stable and accurate than a single tree."

Q: What is gradient boosting?
A: "It builds trees one by one, where each new tree tries to fix the mistakes of the previous one, often giving better accuracy than Random Forest."

Q: What is feature importance?
A: "It shows which features are most useful for making predictions."

3 Data / Preprocessing

Q: Why do we scale features?
A: "Scaling makes features comparable and helps some models (like SVM or KNN) perform better. Random Forest usually doesn't need scaling."

Q: How do you handle missing values?
A: "You can fill them with median, mean, or mode, or remove the missing rows if it doesn't affect much data."

Q: What is one-hot encoding?
A: "It turns categorical data into numbers, with a column for each category marked as 0 or 1."

Q: What is feature engineering?
A: "It's creating new features or transforming data so the model can learn patterns more effectively."

4 Model Evaluation

Q: How do you know your model is good?
A: "We test it on data it hasn't seen and measure errors using metrics like RMSE, MAE, or accuracy."

Q: What is cross-validation?
A: "We split the data into multiple parts, train on some, test on others, and repeat to make sure the model works on all data."

Q: What is bias and variance?
A: “Bias is error from wrong assumptions (underfitting), variance is error from being too sensitive to training data (overfitting).”

5️⃣ Conceptual / Big Picture

Q: How do you choose a machine learning model?
A: “It depends on the data type, problem (classification or regression), amount of data, and complexity needed.”

Q: Why do we split data into train and test sets?
A: “To see how the model performs on unseen data and avoid overfitting.”

Q: Can ML models explain their decisions?
A: “Some do better than others. Random Forest can show feature importance, while deep learning models are usually less interpretable.”