HOME CREDIT DEFAULT RISK

Presented by:
Shania Widianingrum Puspitasari

TABLE OF CONTENT

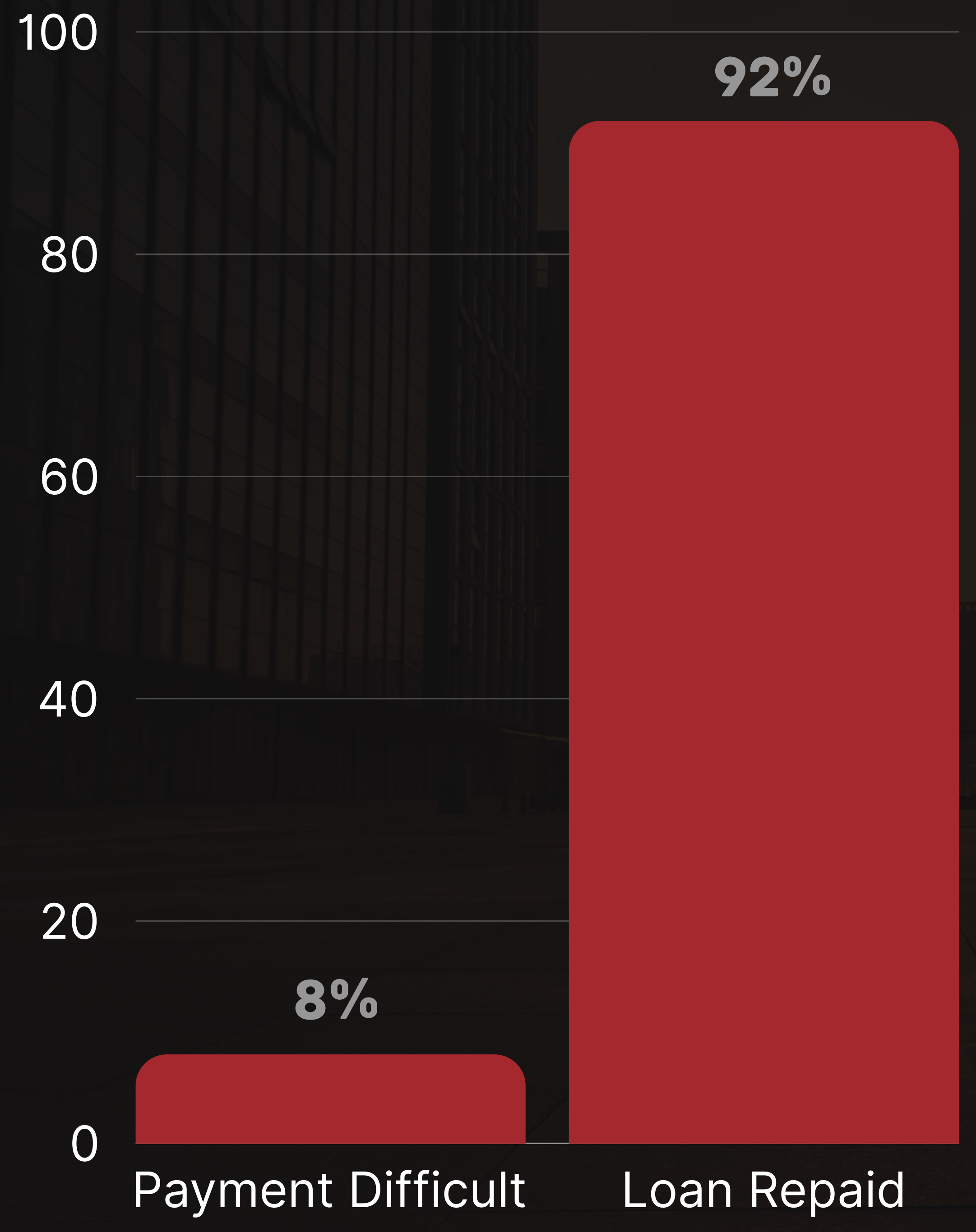
OBJECTIVE

Our main goal is to **create a machine learning model** that can **predict** whether users who will apply for credit can pay on time or will be late / problematic. As a data team, our objective is to **ensure that customers who are able** to make repayments are not rejected when applying for a loan.

METHODOLOGY

1. Data Preprocessing
2. Exploratory Data Analysis
3. Modeling
4. Recommendation

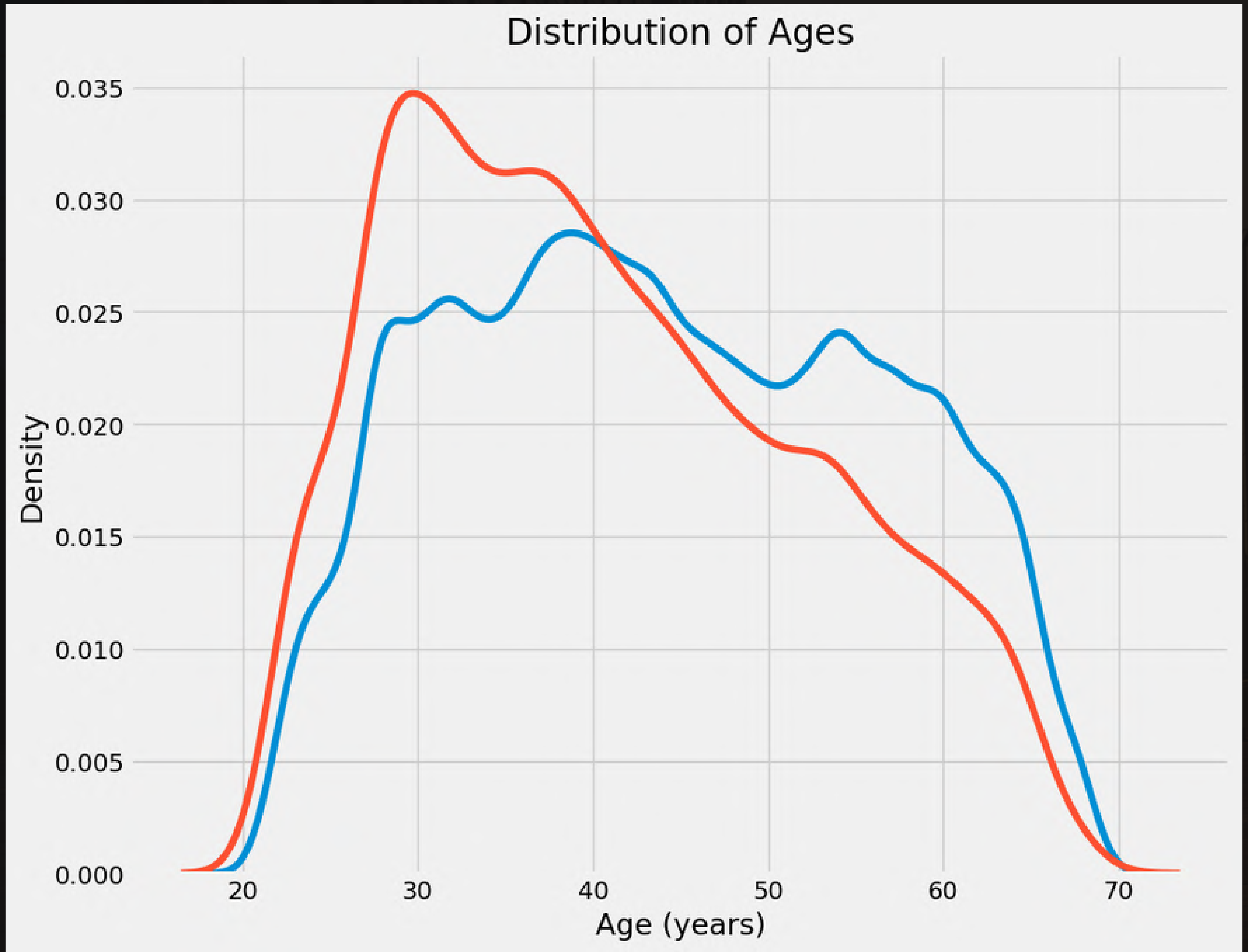
TARGET COLUMN DISTRIBUTION



92%

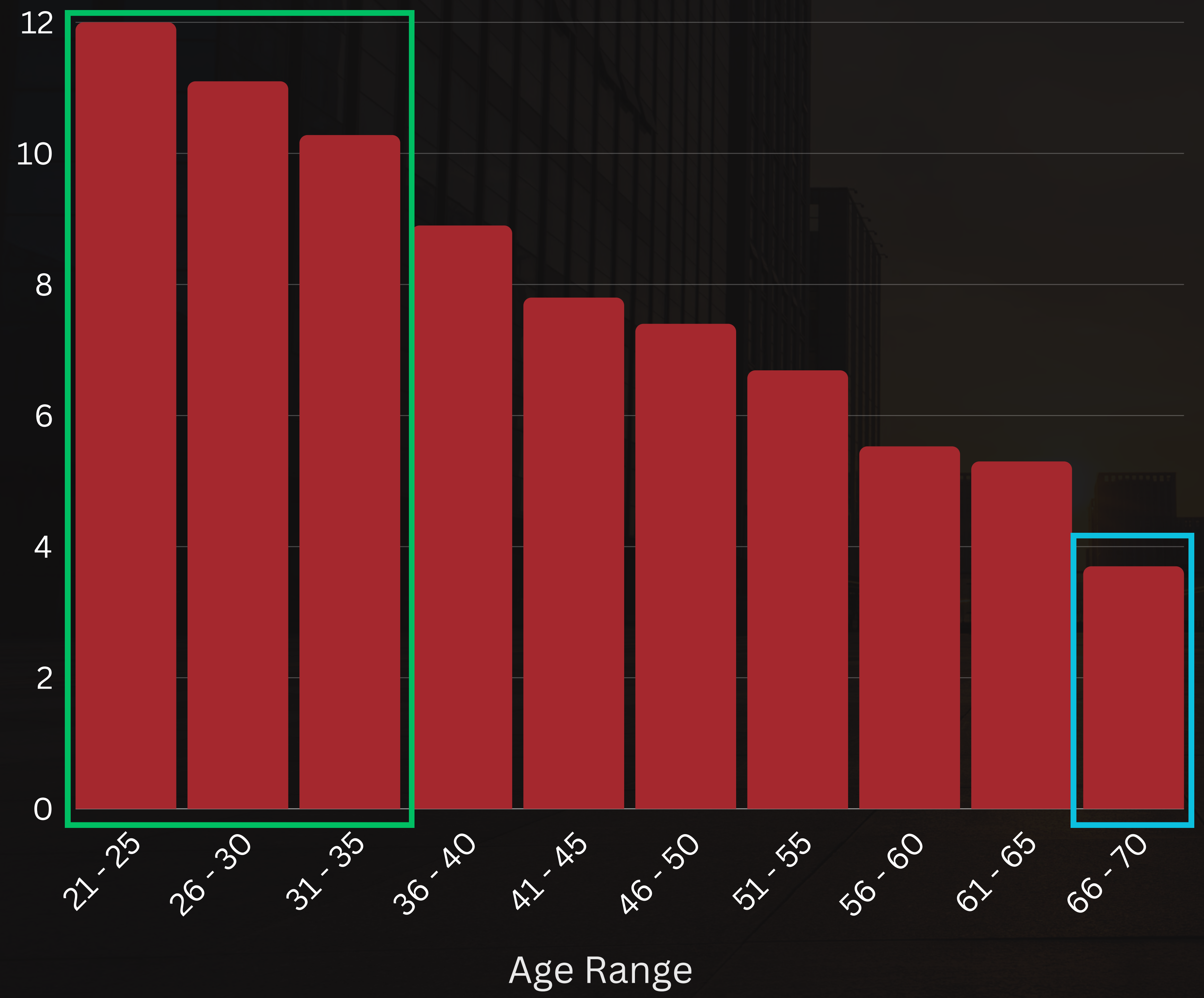
loans are **repaid on time** far more often than defaults.

DISTRIBUTION OF AGES



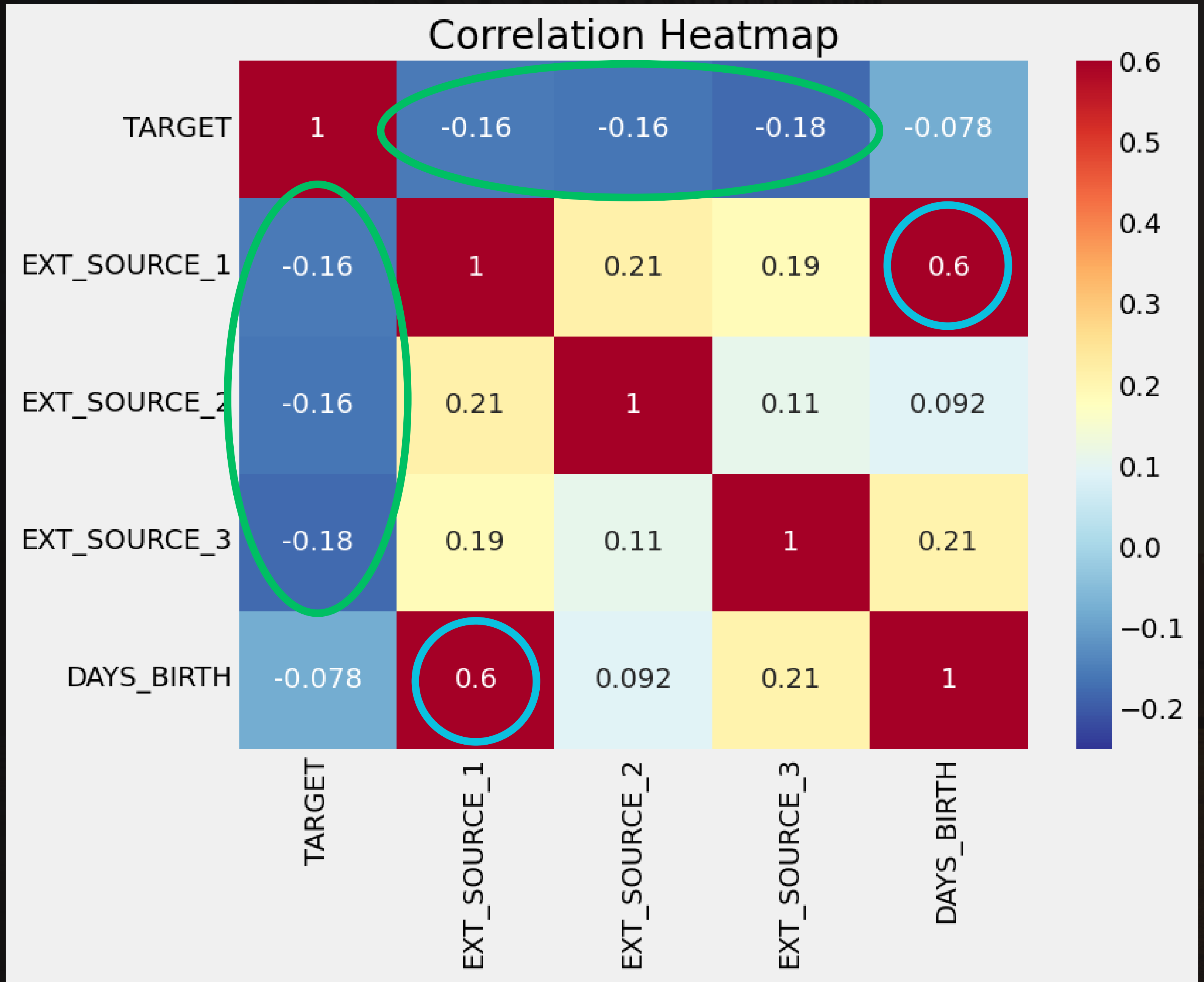
There is a negative linear relationship with the target which means that as **customers age, they tend to repay their loans on time more often**.

EFFECT OF AGE ON REPAYMENT



Younger applicants are **less likely to repay** their loans! Default rates are **above 10%** for the three youngest age groups and **below 5%** for the oldest age group.

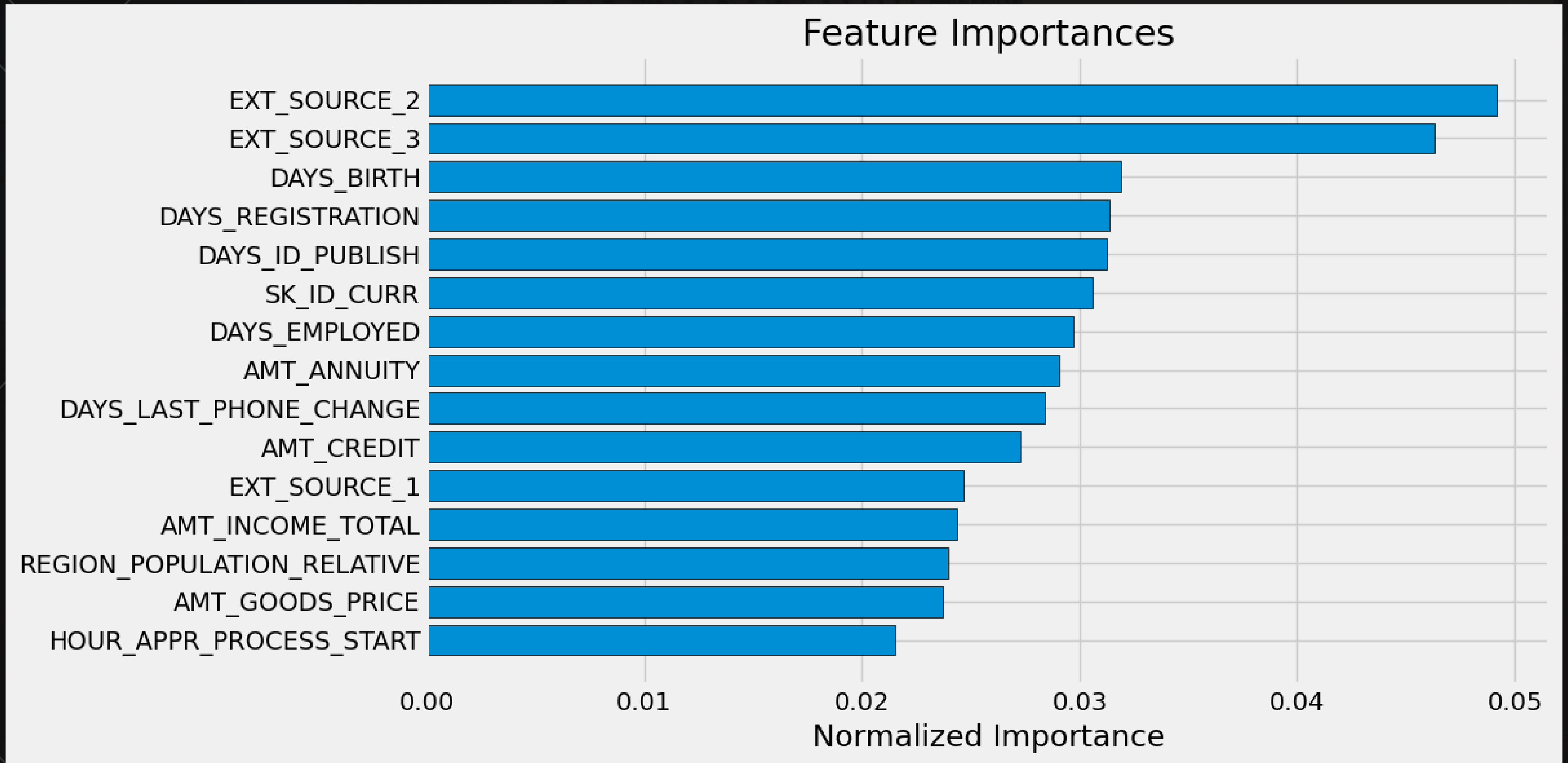
STRONG CORRELATION



All three EXT_SOURCE features have a **negative correlation** with the target, which suggests that as the EXT_SOURCE score increases, it is more likely that the client will repay the loan.

We can also see that DAYS_BIRTH is **positively correlated** with EXT_SOURCE_1 which suggests that perhaps one of the factors in this score is the client's age.

MODELLING AND FEATURE IMPORTANCE



Model	ROC_AUC
Logistic Regression	68
Random Forest	70

As expected, the most important features are those dealing with EXT_SOURCE and DAYS_BIRTH.

We see that all four of our hand-engineered features made it into the top 15 most important! This should give us confidence that our domain knowledge was at least partially on track.

Thank You

Link Project on Github