Capstone Project

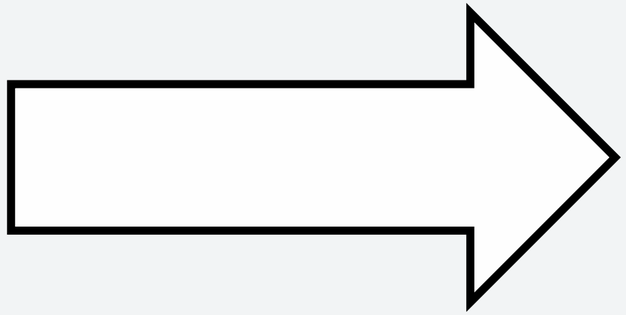
LOAN DEFAULT PREDICTION

Sprint 1

OCTOBER 2023

Presented by:
David Clarke

AGENDA



1. Project Overview
2. The Big Idea
3. The Impact
4. Preliminary EDA
5. Next Steps

PROJECT OVERVIEW

Subject Area:

Finance and Banking Services

Problem Statement:

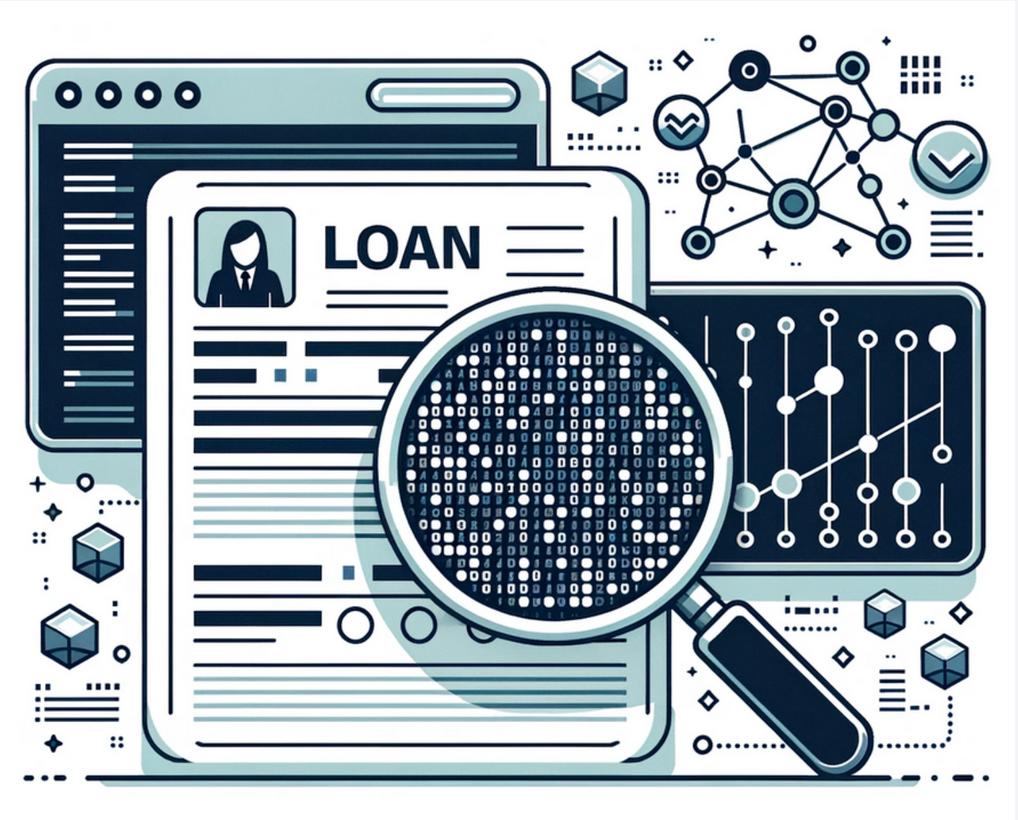
Inability to accurately predict loan defaults:

- financial loss for lending institutions
- reduced trust in financial systems
- decreased credit availability
- financial crises

Opportunity Identified:

Leverage data and ML to predict loan defaults

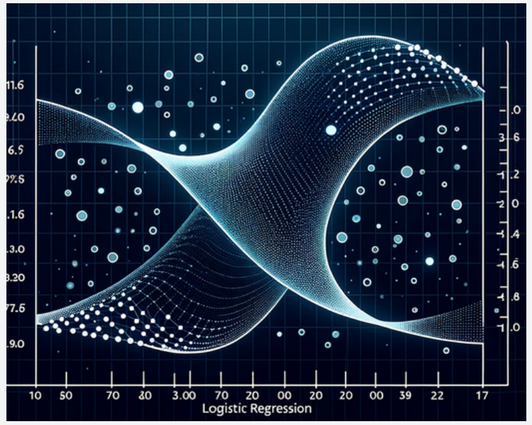
- increase accuracy

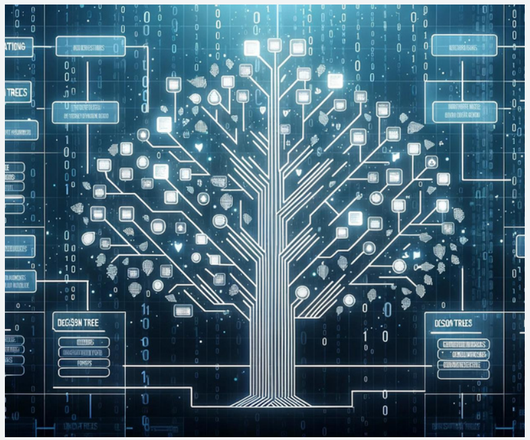


THE BIG IDEA

Objective:

Build a predictive model that can forecast loan defaults based on certain demographic data and financial information.





Data Science Application:

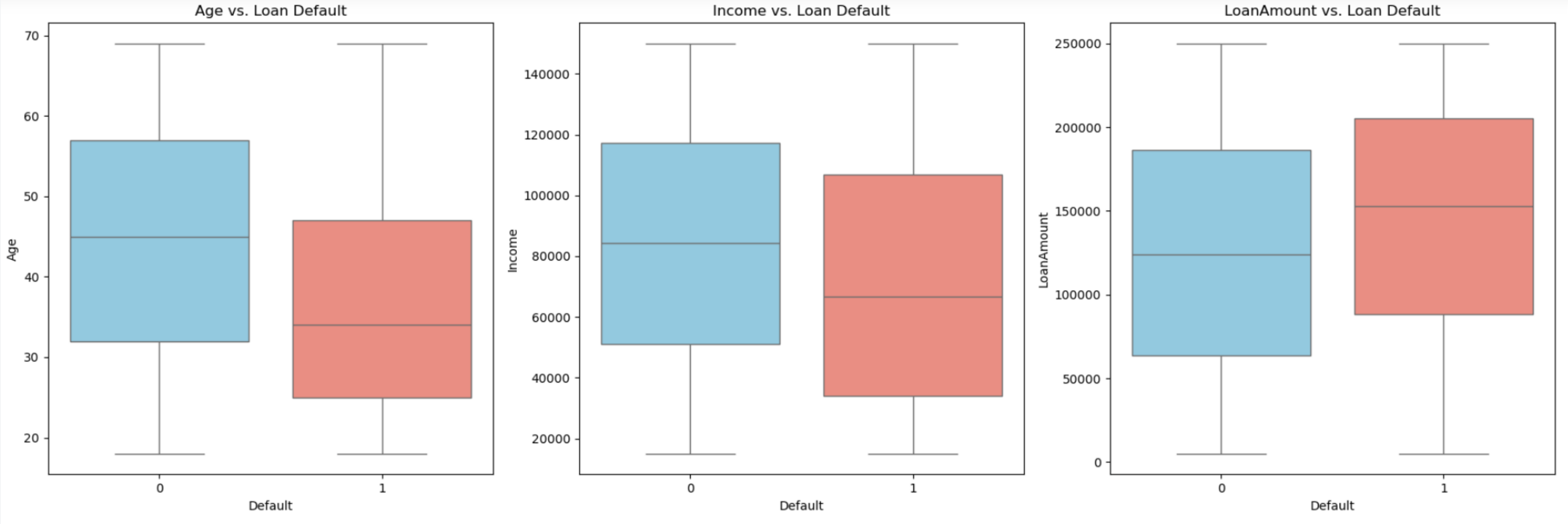
- Use historical loan data to train ML models
- Use advanced algorithms that can capture intricate patterns and relationships in the data
- Continuously refine and improve

THE IMPACT

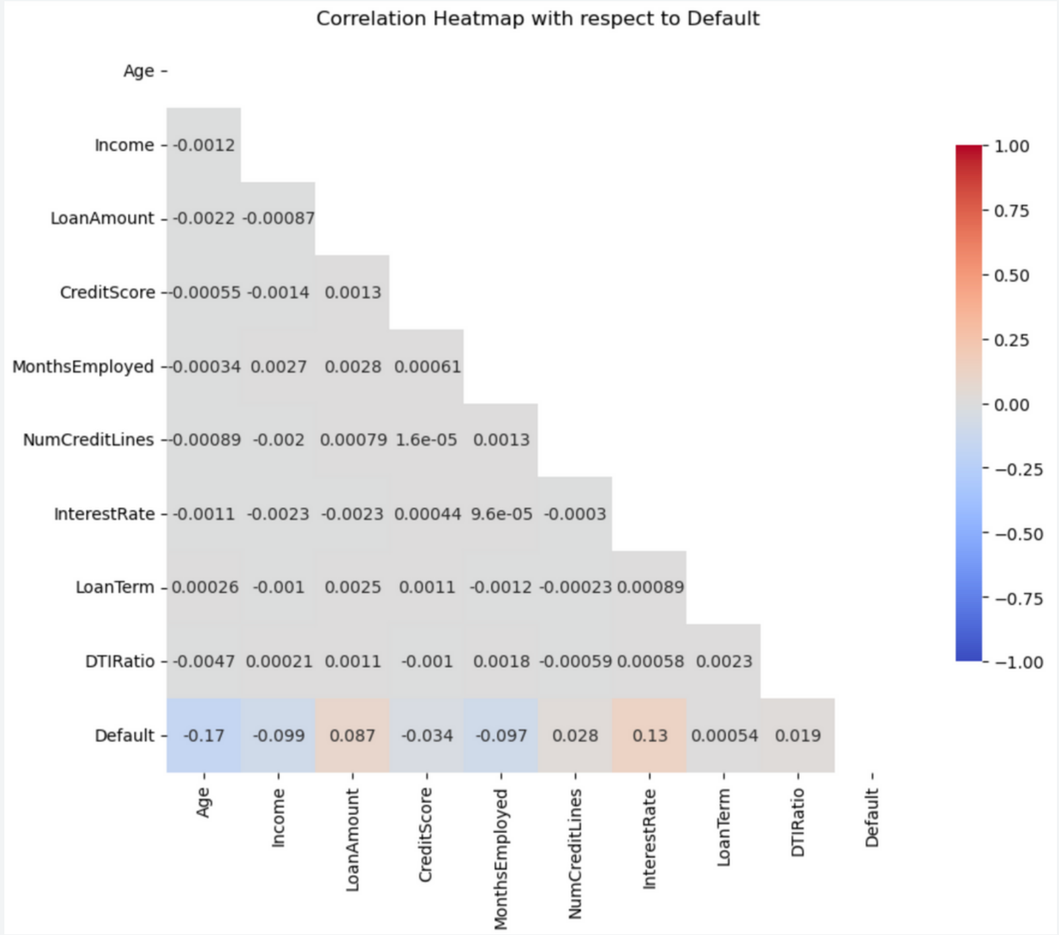
A *very* simplistic approach to quantifying impact:

- Approx. **2.4%** of all consumer loans at commercial banks in the US were reported as delinquent. (*Statista 2023*)
- Approx. **$17 trillion** in total consumer debt in US (*The Fed 2023*)
- Approx. **$400 billion loss** (2.4% * 17T - assuming all delinquent -> defaulted)
- A mere 1% improvement in prediction accuracy could result in approx. $4B in savings
- The project's aim, even by a modest margin, could translate to billions of dollars in savings, given the vast sums involved in lending.

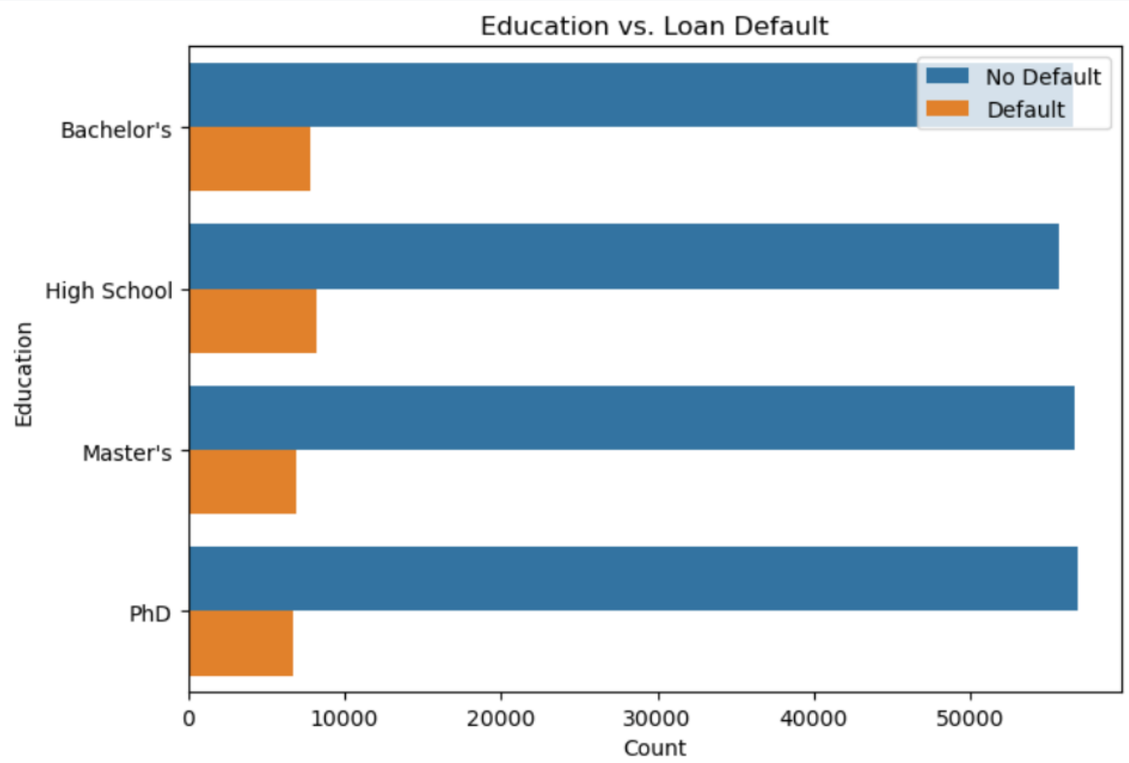
PRELIMINARY EDA

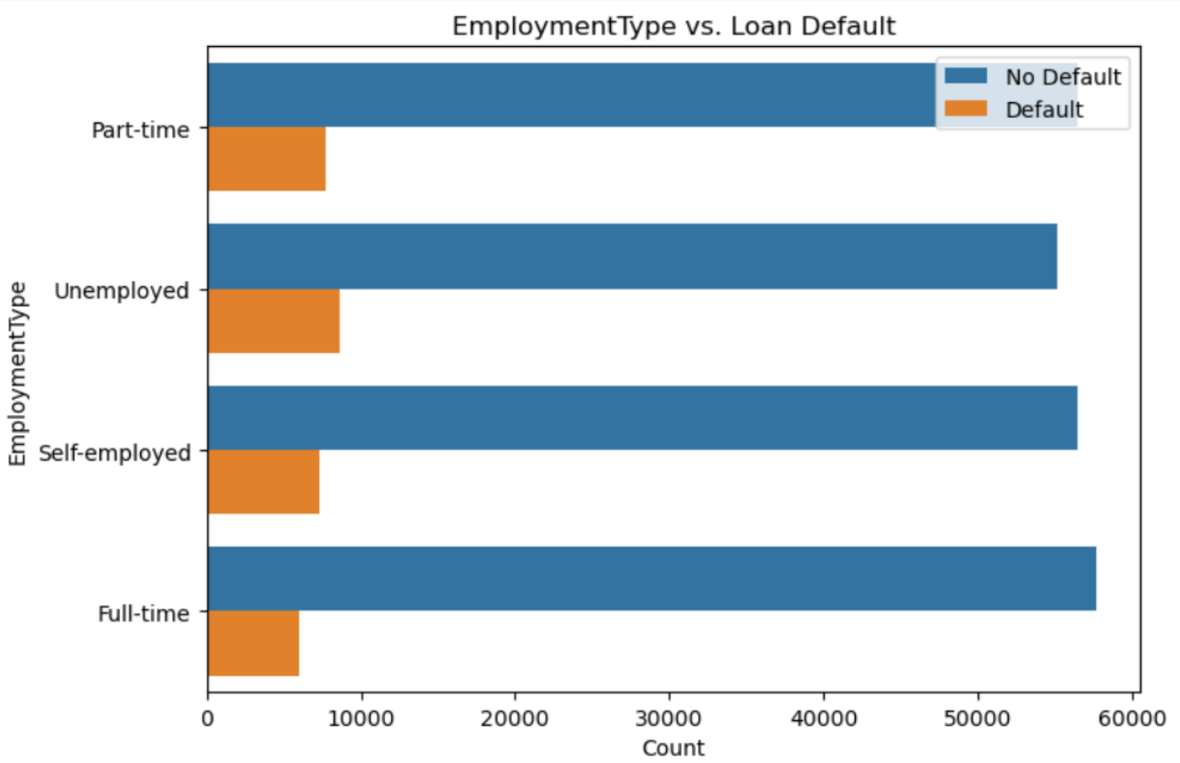


PRELIMINARY EDA



PRELIMINARY EDA





NEXT STEPS

Data preprocessing techniques:

- **One-hot Encoding** as categorical features like Education, EmploymentType and MaritalStatus need to be converted to numeric values which would be suitable for machine learning.

- **Feature Scaling** in order to normalize certain features to bring them to a similar scale. Features like Income, LoanAmount, and CreditScorehave different scales.

Feature engineering opportunities:

- **Binning Continuous Variables**
  - Age -> "Young", "Middle-aged", "Senior".
  - Income levels -> "Low", "Medium", "High".
  - CreditScore -> "Poor", "Good", "Excellent".

- **Creating new features** based on interactions between existing features within the dataset

Thank you