

## Capstone Session 6

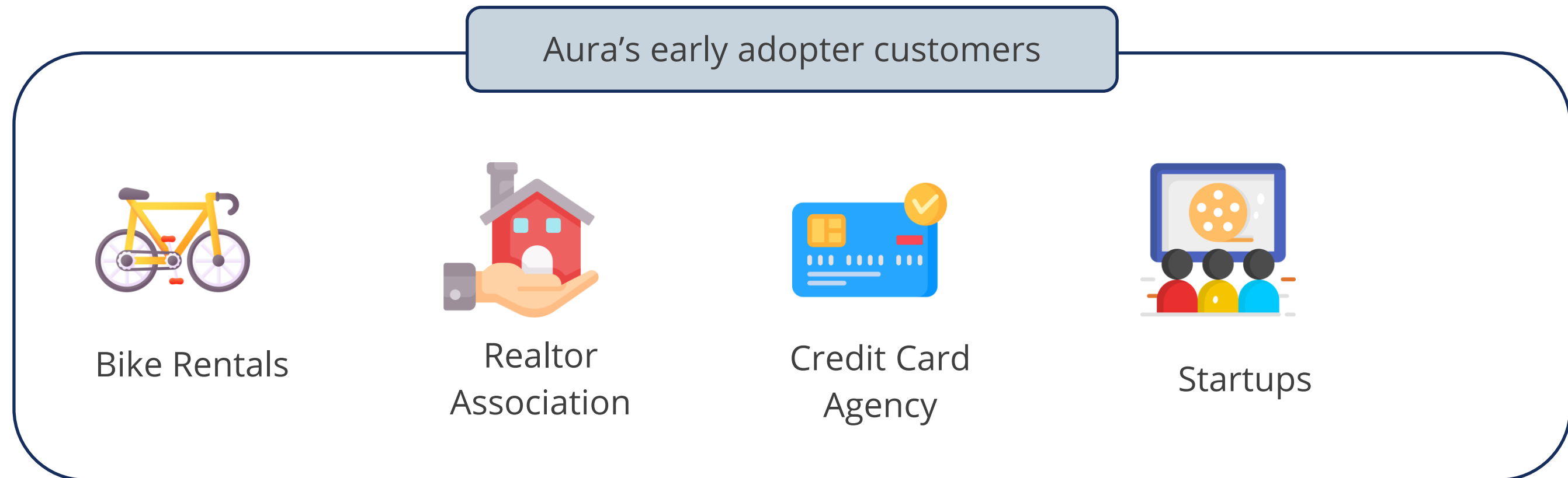




# **Machine Learning for Modeling**

# Machine Learning for Strategic Decision-Making

The intelligence provided by Aura will help customers make decisions for their omnichannel marketing and customer acquisition programs.



# Project Statement

Aura must do the following:



Predict bike-sharing demand

Classify incomes

Cluster credit card users

Build a recommendation engine

# Session 6 : Dataset Description

Adultcensusincome.csv

Variable	Description	Variable	Description
Age	Age of the person	workclass	Workclass of the person
fnlwgt	Weighted tally of specified socio-economic characteristics of the population	Education	Education level of the person
education.num	Number of years of education	marital.status	Marital status of the person
occupation	Occupation of the person	relationship	Relationship status of the person
capital.gain	profit for assets	sex	The person's sex (Male/ Female)
hours.per.week	Number of working hours per week	native.country	The native country of the person
Income	The income category of the person	capital.loss	loss for assets

## Session 6: Building an Income Classification Model

---

**Task:** Build a classification model for predicting the income using the Adult Census Income Dataset.

- Load the dataset
- Check for null values and ? in any columns and handle those values. Check the distribution of target variable income and identify if the dataset is balanced.
- Perform the following Univariate analysis
- Create a barplot for column income
- Create a distribution plot for column age
- Create a barplot for column education
- Create a barplot for Years of Education. Use column education.num
- Create a pie chart for Marital status. Use column marital.status

## Session 6: Building an Income Classification Model

---

- Perform the following Bivariate analysis
  - Create a countplot of income across columns age, education, marital status, sex
  - Draw a heatmap of data correlation and find out the columns to which income is highly correlated
- Prepare the dataset for modeling
  - Label encode all the categorical columns
  - Prepare independent variables X and dependent variable Y (Income).
  - Perform feature scaling using StandardScaler and fix the imbalance in the dataset using any one of the techniques like SMOTE or RandomOverSampler
  - Perform a train test split in the ratio 80:20 and random\_state 42.
- Perform Data Modeling
  - Train Logistic Regression Model, KNN Classifier Model, SVM Classifier, Naive Bayes Classifier, Decision Tree Classifier and Random Forest Classifier
  - Perform model evaluation on Accuracy and F1 score and identify the best model



**Thank You**