DATA AND
ARTIFICIAL INTELLIGENCE

Capstone Session 6

simplilearn

DATA AND
ARTIFICIAL INTELLIGENCE

Machine Learning for Modeling

simplilearn

Machine Learning End Goal

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

Aura's early adopter customers

Bike Rentals

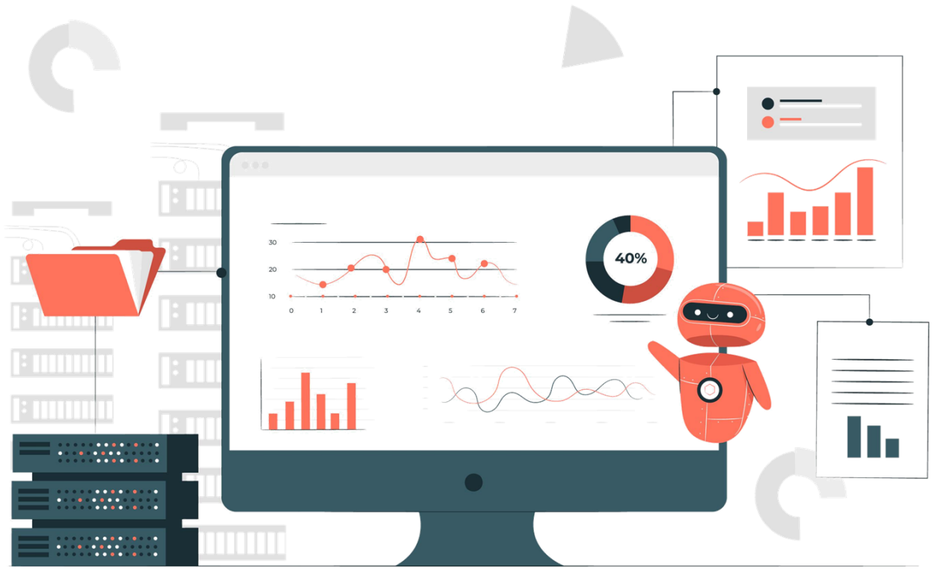
Realtor Association

Credit Card Agency

Startups

Project Statement

Aura must do the following:



- Predict bike-sharing demand
- Classify incomes
- Cluster credit card users
- Build a recommendation engine

Week 6: Dataset Description

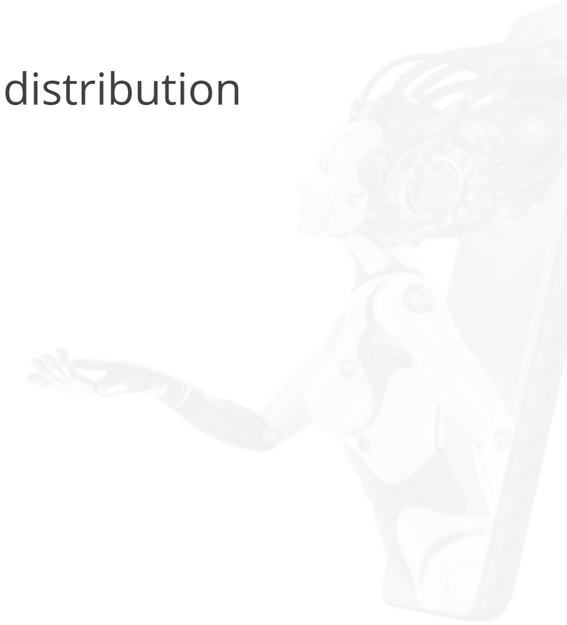
Adultcensusincome.csv

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
race	The race of the person	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	-	-

Week 6

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 below 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



Week 6

- Perform below Bivariate analysis
  - Create a countplot of income across columns age, education, Marital Status, race, 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

DATA AND
ARTIFICIAL INTELLIGENCE

Thank You

simplilearn