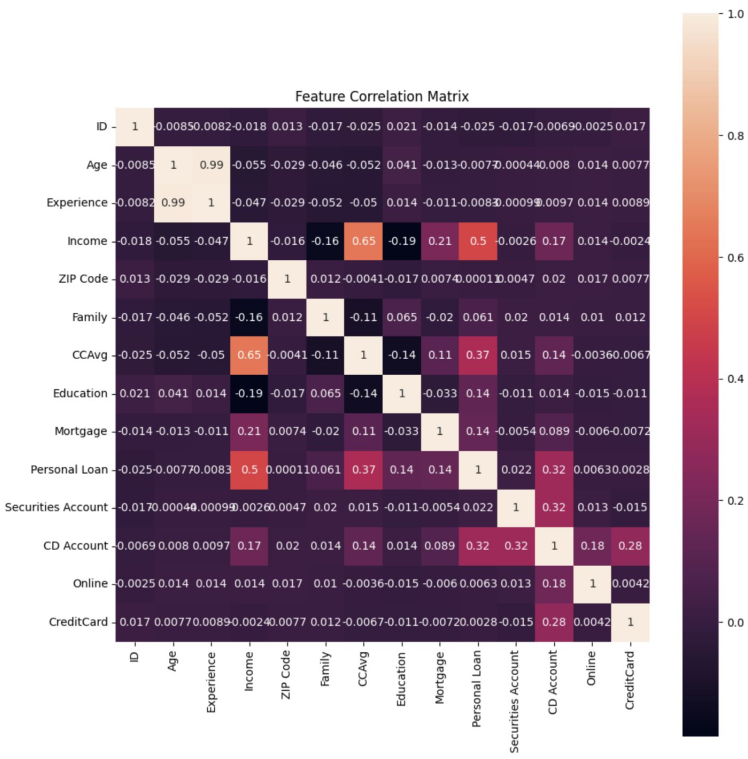
PERSONAL BANKING LOAN PREDICTION

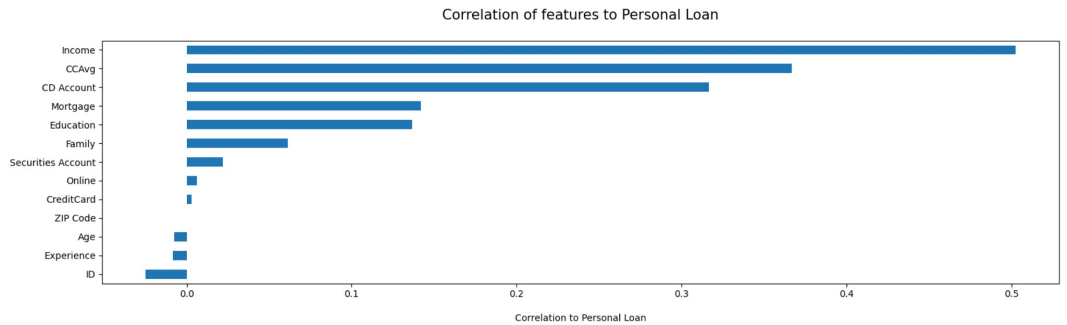
W4995 Applied Machine Learning

Thursday Section Group 18

Yizhi Liu (yl4993), Ji Qi (jq2365), Wen Song (ws2685), Unal Yigit Ozulku (uyo2000)

SECTION I: INITIAL DATA EXPLORATION





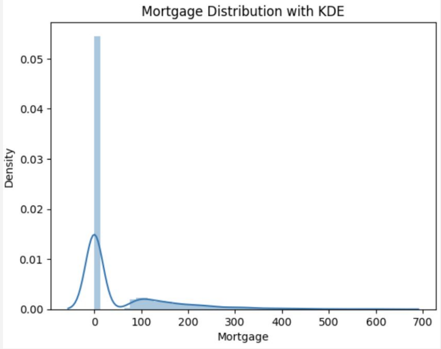
ANALYSIS:

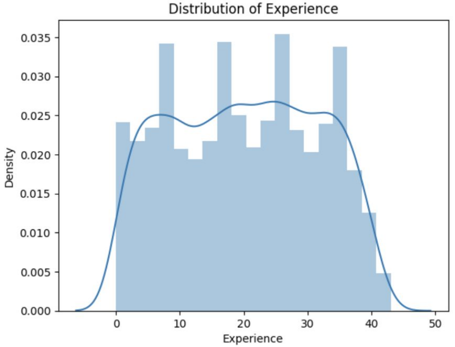
According to the correlation heatmap and the bar graph we plot above, we have the following initial observations:

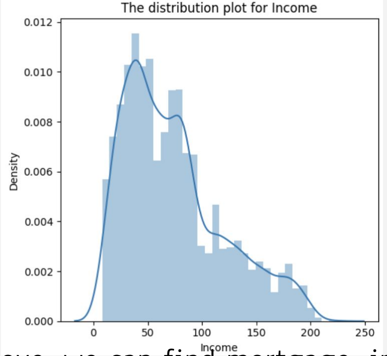
- The top 3 most important features to Personal_Loan: "Income", "CCAvg", and "CD Account"
- From the correlation heatmap, we can find "Age" and "Experience" are highly correlated and almost linearly related to each other, with correlation 0.99 with each other.

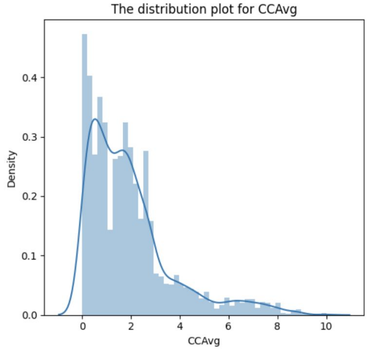
UNIVARIATE ANALYSIS

NUMERICAL FEATURES



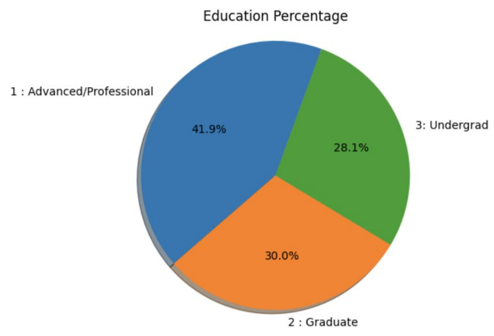


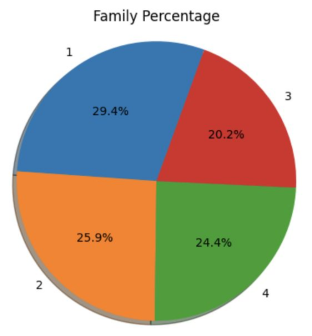


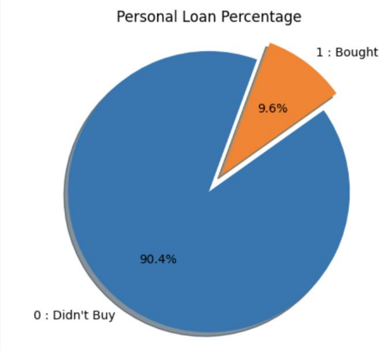


According to the distribution graphs for the 4 numerical features above, we can find mortgage, income and CCAvg all skewed to the right, meaning that we need to use normalization techniques to process data to avoid distortion in our output.

CATEGORICAL FEATURES



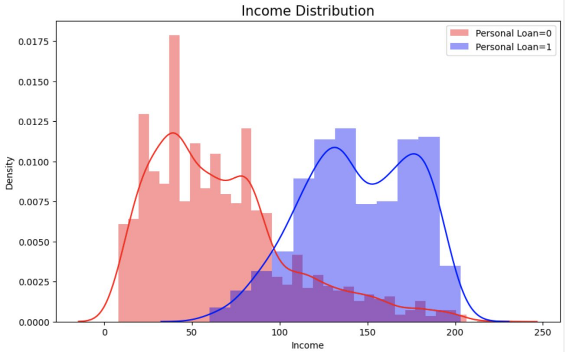


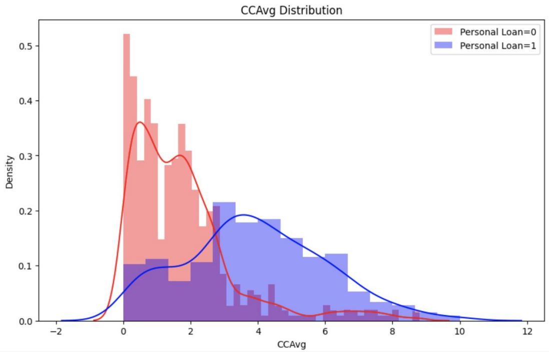


According to the pie chart on the left, we can find out of all customers, only 9.6% customers bought loan in the last campaign. So data imbalance issues exist in our dataset. We need to deal with this imbalance issue in the later part.

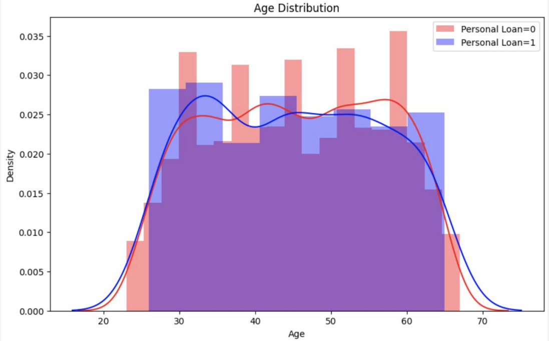
Education - 42% of candidates have bachelor's degree and 30% have master's degree and 28% are professionals.
Family - Around 29% of the customer's family size is 1, 26% is 2, 20% is 3 and 24% is 4.

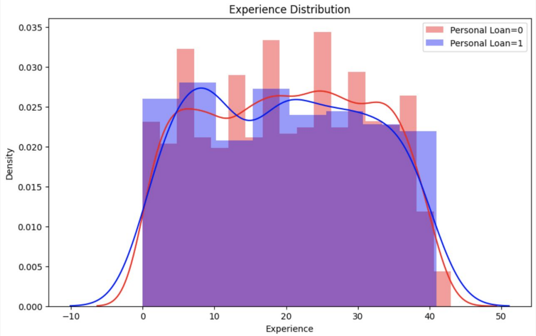
BIVARIATE ANALYSIS - DIFFERENT FEATURES VS. PERSONAL LOAN AND RE



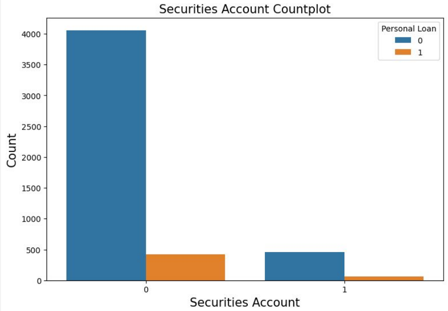


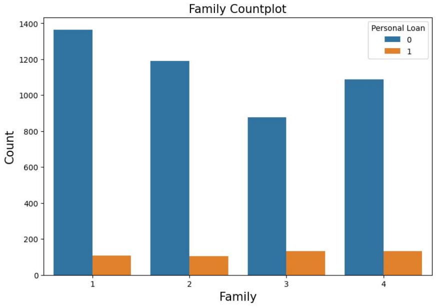
For the **numerical features**, we mainly use **distribution graph** to explore the relationships between these features and our target variable.
According to the distribution graph of Income VS. Personal Loan and CCAvg VS. Personal Loan, we can find people who has personal loan usually have higher income. For people whose income is lower, they are less likely to have personal loan; People who has personal loan usually have higher monthly spending with the credit card. For people whose income is lower, they spend less with the credit card each month.

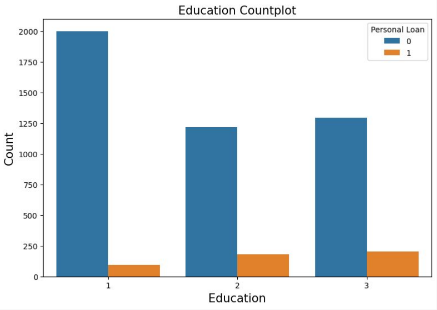




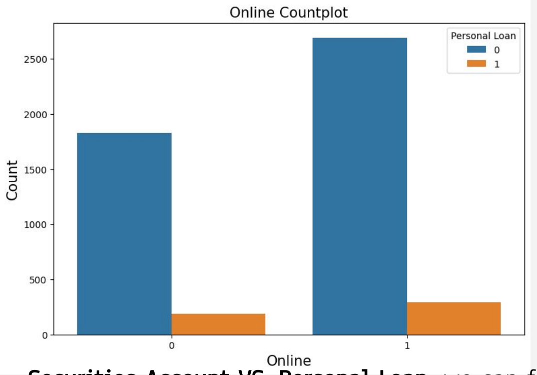
Based on our previous conclusion, we know age and year of experience are closed positively related to each other. So their distribution graphs show roughly the same pattern. Moreover, the distribution are kind of average for these two features vs. personal loan. Thus, these two features might not influence our target variable too much.

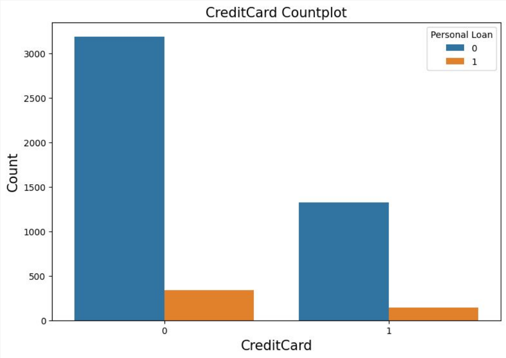


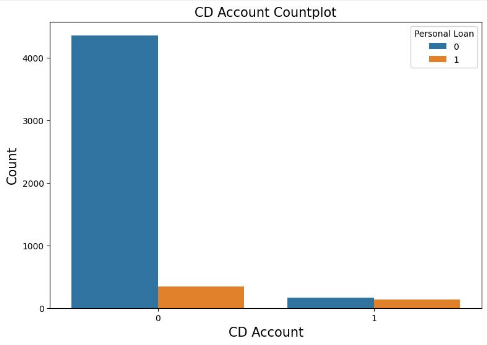




For the **categorical features**, we mainly use **countplot** to explore their relationship with our target variable - personal loan. As the three countplots shown on the lefts side. The remaining countplots are shown on the next slide.







Securities Account VS. Personal Loan: we can find most people don't have securities account. Moreover, out of which, most people don't have personal loans either.

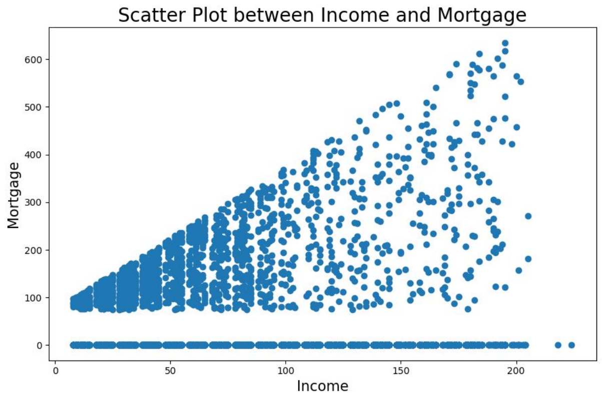
Family VS. Personal Loan: For person who do not have personal loan, their family size tend to be 1, followed by 2. This is a reasonable observation since their expense might be less than other categories of customers. Then for people who have personal loan, they usually have a larger family size.

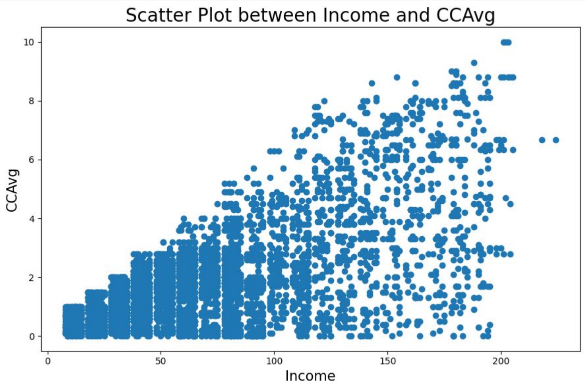
Education VS. Personal Loan: Customers with lower educational qualification are more likely to buy personal loan. As they should have a lower income. But customers who have high qualified education background are less prone to buy personal loan.

Online VS. Personal Loan: Most customers has online internet banking facility and the people who have these facilities have bought more number of personal loans.

CreditCard VS. Personal Loan: Most customers don't have a credit card. Moreover, the maximum number of customers who bought personal loan don't have credit card, this might because they don't have the advantages of using credit card. So they are prone to use personal loan.

CD Account VS. Personal Loan: Most customers don't have Certificate of deposit account. Moreover, the maximum number of customers who bought personal loan also don't have CD Account.





From the scatter plots on the left, we can find many of the customers don't have any mortgage in their name. So we can observe a line along the mortgage value zero. And the mortgage value for the customers often increases with increase in their income. Customers with high income are more likely to have higher CCAvg and more likely to buy Personal Loan as what we observed before.

SECTION II: CLEANING AND PREPROCESSING

DATA CLEANING

1. After the initial data exploration, we found that the dataset **doesn't contain any missing** or **duplicate values.**

2. Based on the Univariate Analysis, we need to **drop all the noise records whose 'Experience' <0 or 'ZIP Code' < 5 digits.**

3. **Drop The variable ID**, since it only acts as a customer identifier without adding any useful information to the dataset.

4. **Drop 'Experience'** (strongly correlated with Age.)

FEATURE ENGINEERING

1. **Log transformation** will be applied to remove the right skewness in **'Income', "CCAvg', and 'Mortgage'**(too many zeros and plus one before log transformation).

2. **Convert 'CCAvg' average monthly credit card spending to annual spending** by times 12 (Equal the unit in the Income)

3. **Category Encoding** will be selected to convert **'Family', 'Education' and 'Zip Code'** into numerical values.

ADDRESS IMBALANCED DATA

1. **Target Variable: Personal Loan** (Whether the potential customers have a higher probability of purchasing the loan)

2. **Only 9.6%** customers bought loan in the last campaign

3. **Synthetic Minority Oversampling Technique & Train-Test Split (80:20)-** Reduce Overfitting incurred by Oversampling

4. **After All the above steps, the cleaned dataset has 12 columns and 4881 rows.**

SECTION III: INSIGHTS FROM DATA EXPLORATION

SUMMARY:

From the detailed insights in previous slides, we can get the following summary:

I. From correlation matrix/heatmap, we gained the insights
   A. the most 3 relevant features to our target variable (Personal Loan) are "Income", "CCAvg", and "CD Account".
   B. High multicollinearity exists between “Age” and “Experience”, so we might need to drop any one of them to avoid some distortions in our output.

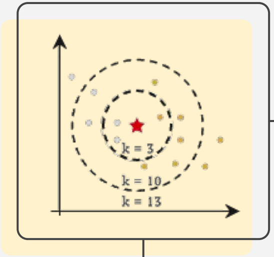
II. Speaking of the numerical features in our dataset, the distribution graphs for “Mortgage”, “Income” and “CCAvg” all have skewed distributions, so we need to take normalization techniques and encoding methods (such as log transformation, target encoding, etc) to solve potential bias/distortion problems.

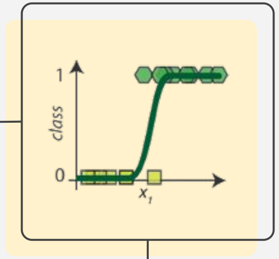
III. Notice data imbalance issue exists when analyzing the pie chart of Personal Loan, we can use machine learning techniques (such as SMOTE, F1-Score, etc.) to relieve the issues induced by data imbalance.

MACHINE LEARNING TECHNIQUES PROPOSED TO BE IMPLEMENTED

K-NEAREST NEIGHBOR

KNN is a non-parametric technique for classification that can be used to predict whether a customer is likely to take a personal loan or not based on the similarity (distance) of the features



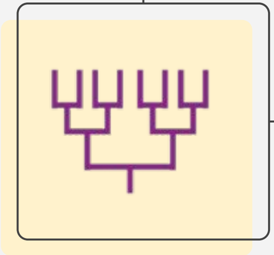


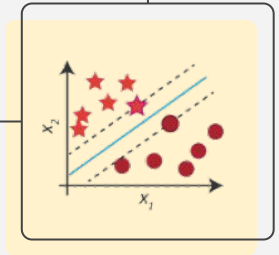
LOGISTIC REGRESSION

Logistic regression is a widely accepted method for the binary classification problem (i.e., whether a customer is likely to take a personal loan or not). This method can be enhanced by using ***kernel trick***.

DECISION TREE

Decision tree classifier is easy to interpret and visualize, and can handle the ***categorical*** and ***numerical*** features included in the Personal Loan Dataset

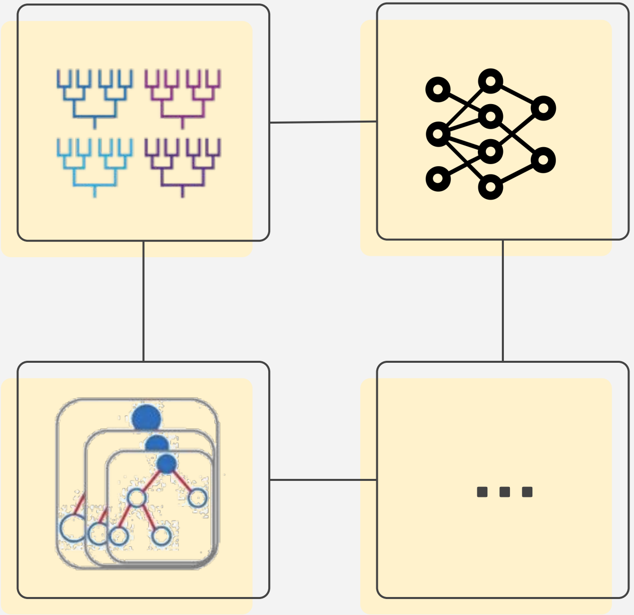




SUPPORT VECTOR MACHINE

SVM can be an efficient classifier for binary classification problem, it can handle the linear and non-linear decision boundaries using the ***kernel trick*** (suitable for the ***data processing*** step). The kernel trick can increase the accuracy and robustness of SVM

MACHINE LEARNING TECHNIQUES PROPOSED TO BE IMPLEMENTED



RANDOM FOREST

This is an ensemble method that combines multiple decision trees to improve the classification. It can also *provide feature importance rankings*

NEURAL NETWORK

Neural network is a powerful class of machine learning models that can be applied for the Personal Loan Dataset. Compared to other classifiers, it can capture extract patterns from features and detect the subtle interactions between features

GRADIENT BOOSTING

Gradient boosting is also an ensemble method that combines multiple weak classifiers (e.g., decision tree) to create a strong predictor. This method can handle both categorical and numerical features, and *provide feature importance rankings*

FURTHER STEPS

Based on our further exploration and feature engineering steps for the dataset, we will test additional classifiers to identify the most effective model for predicting load acceptance. We will consider more advanced techniques, such as the advanced neural networks we will learn from the lecture

CLASSIFICATION PERFORMANCE

The Personal Bank Loan dataset is an ***Imbalanced Dataset*** → ***F1-score as a metric***

CLASSIFIER	F1-SCORE
K-NN	79.36%
DECISION TREE	90.91%
LOGISTIC REGRESSION	69.79%
SUPPORT VECTOR MACHINE	91.12%

CLASSIFIER	F1-SCORE
RANDOM FOREST	93.98%
GRADIENT BOOSTING	95.16%
NEURAL NETWORK	*96.29%*
WILL BE UPDATED IN NEXT STEP	

We will also use other metrics to evaluate the performance of the ML classifiers for the Personal Bank Loan Dataset, including ***Accuracy, Precision, and AUC***