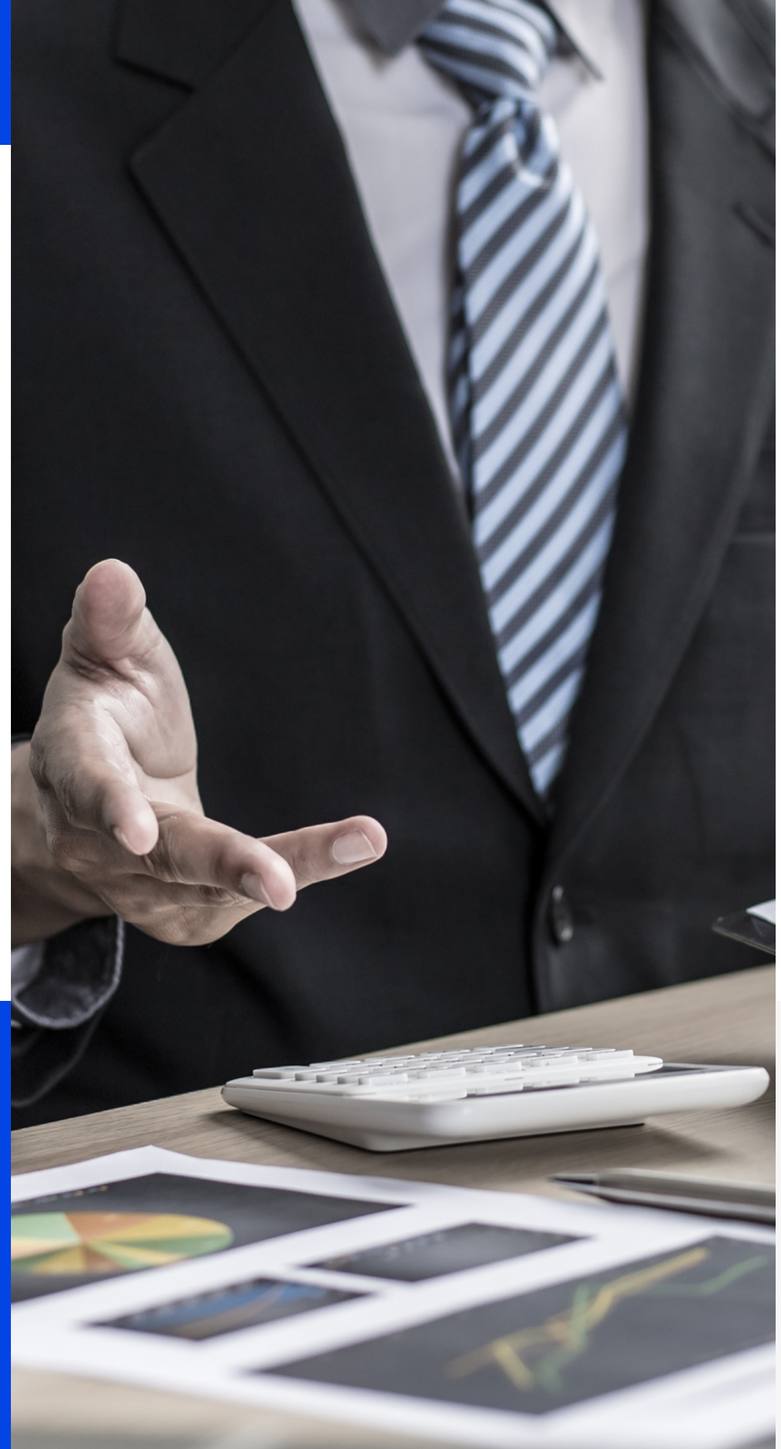
LOAN DEFAULT PREDICTION

Group 2:



CONTENT

LOANS UNDERSTANDING

LOAN DEFAULT STATISTICS

- **Global economic impact**:
  - In 2023, Global debt reached **$235** trillion; **238%** of GDP - International Monetary Fund
- **Mortgage Defaults and Housing Market**:
  - The **2008 financial crisis**, triggered in part by widespread mortgage defaults, had severe consequences on the global economy.
  - Between 2006 and 2014, the median net worth of American families fell by **40%**, emphasizing the need for effective loan default prediction to avoid similar crises in the future - U.S. Federal Reserve
- **Student Loan Defaults**:
  - In 2022, outstanding student loan debt in the U.S. reached **$1.76** trillion - U.S. Federal Reserve

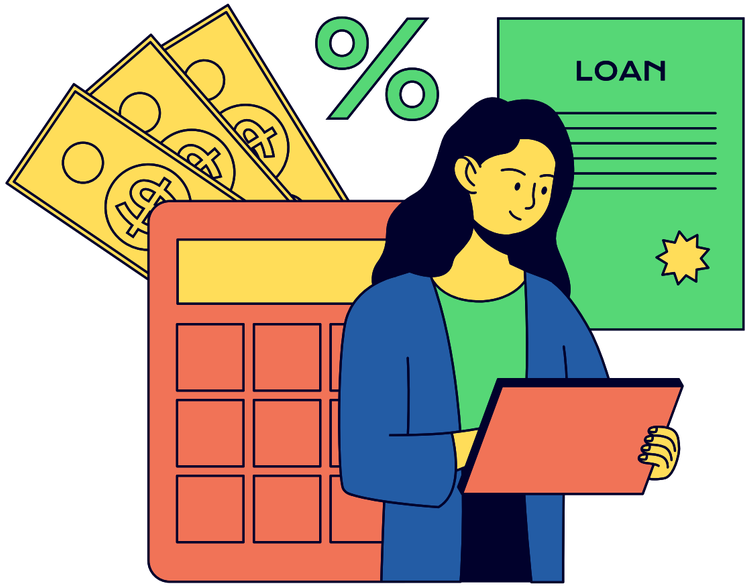
WHY DEFAULT PREDICTION IS IMPORTANT

- Primary objective of banks and financial institutions is to reduce payment default and ensure borrowers repay their loans as expected.
- Defaulted loans can result in financial losses of the lender.
- Predictive models can help identify at-risk borrowers early, reducing default rates and improving loan portfolio performance.

DATA SOURCE

DATA SOURCE

- Loan Default Prediction Dataset
  - Found on Kaggle
  - Taken from Coursera’s Loan Default Prediction Challenge
  - The dataset contains 255,347 rows and 18 columns in total.
- Contains important columns that can help likelihood to default on a loan
  - Education, Employment Type, Credit Score, Income, etc.
- Experian.com



DATA DICTIONARY

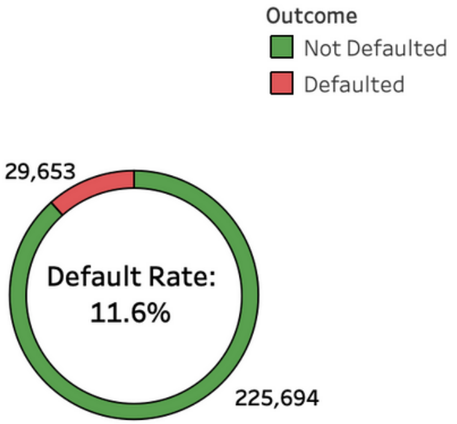
- **Default**: Outcome variable; Indicates whether the loan defaulted or not
- **LoanID:** A unique identifier for each loan; 255,347 unique values
- **Age**: The age of the borrower
- **Income**: The annual income of the borrower
- **LoanAmount**:The amount of money being borrowed
- **CreditScore**: The credit score of the borrower
- **MonthsEmployed**:The number of months the borrower has been employed
- **NumCreditLines**:The number of credit lines the borrower has open
- **InterestRate**:The interest rate for the loan
- **LoanTerm**:The term length of the loan in months
- **DTIRatio**:The Debt-to-Income ratio
- **Education**: The highest level of education attained by the borrower
- **EmploymentType**: The type of employment status of the borrower
- **MaritalStatus**: The marital status of the borrower
- **HasMortgage**: Whether the borrower has a mortgage
- **HasDependents**: Whether the borrower has dependents
- **LoanPurpose**: The purpose of the loan
- **HasCosigner**: Whether the loan has a co-signer

DATA UNDERSTATING

LOAN SUMMARY STATS

- Average of defaulted loan amount > not defaulted

Analysis of Loan's

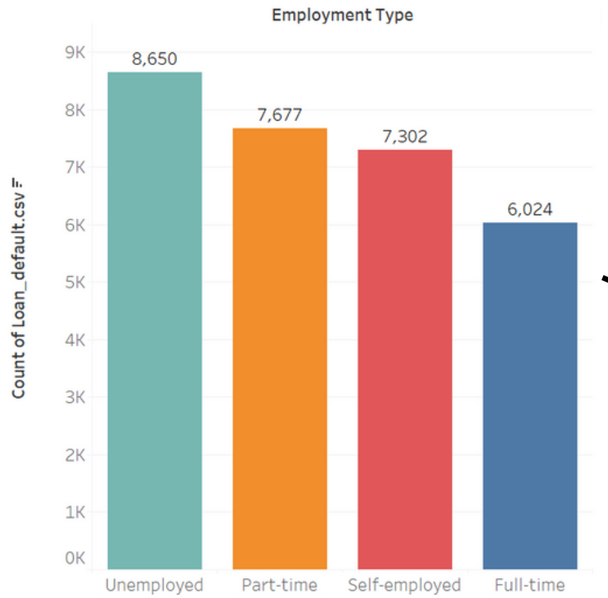


Defaulted Loan Breakdown (Averages)

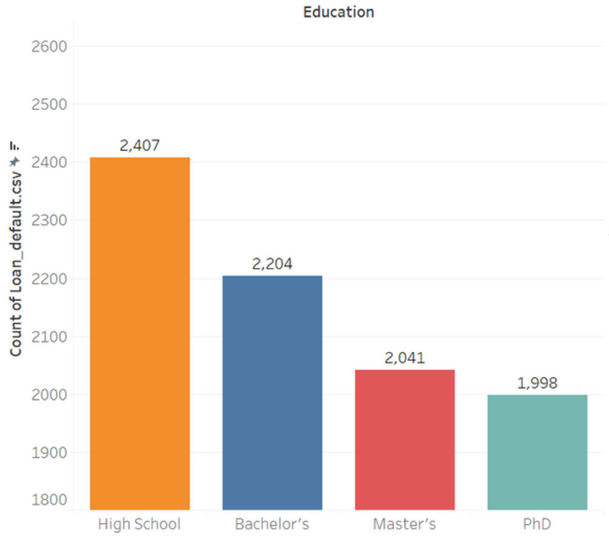
	Outcome	
	Not Defaulted	Defaulted
Loan Amount	$125,353.7	$144,515.3
Loan Term	36.0	36.1
DTI Ratio	0.499	0.512
Age	44.4	36.6
Credit Score	576.2	559.3
Income	$83,899	$71,845
Interest Rate	13.2	15.9
Months Employed	60.8	50.2
Number of Credit Lines	2.5	2.6

PROFILING A LOAN DEFAULTER

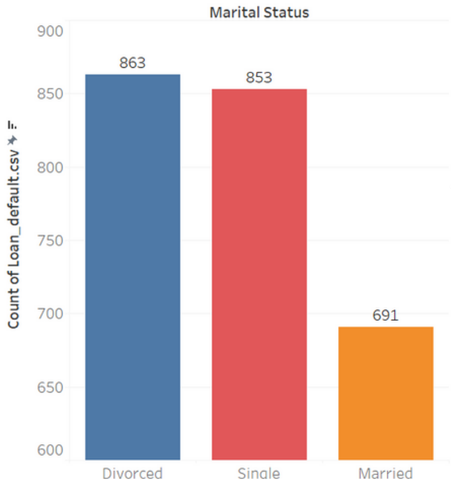
Top Defaulters



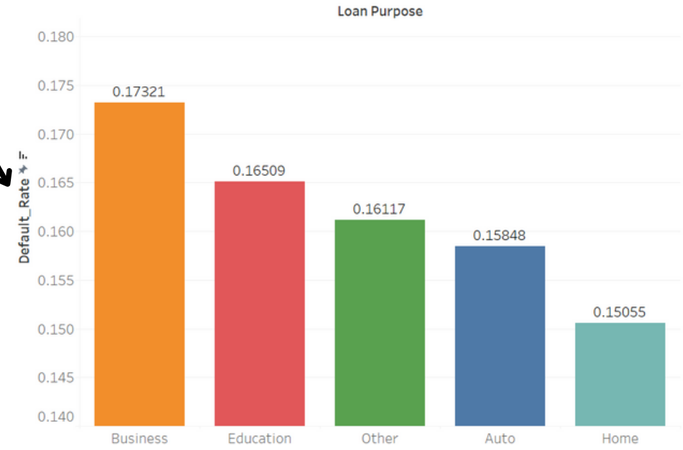
Does education Level affect Loan Defaulting Amongst Unemployed Folks?



Does Marital Status affect Probability of Default for Unemployed, High School Grads?



Which Loan Purpose has the highest default rate Among Unemployed Divorcees who are High School Graduates



Through the profiling, we realize that

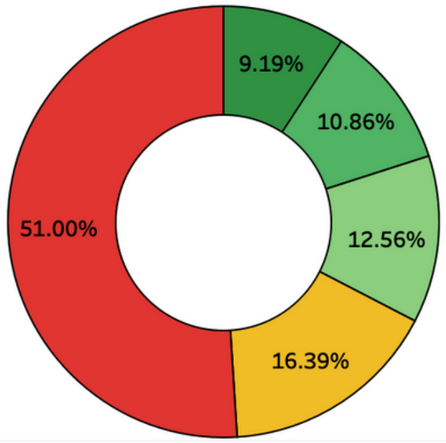
1. Unemployed Folks are the highest Loan Defaulters
2. Among Unemployed customers, those who are only high school graduates have high chances to default
3. Divorcees who are unemployed high school grads are high risk customers
4. Business Loans, followed by Education Loans are susceptible to default if provided to unemployed divorcees with a high school degree

CREDIT SCORE GROUPS BREAKDOWN

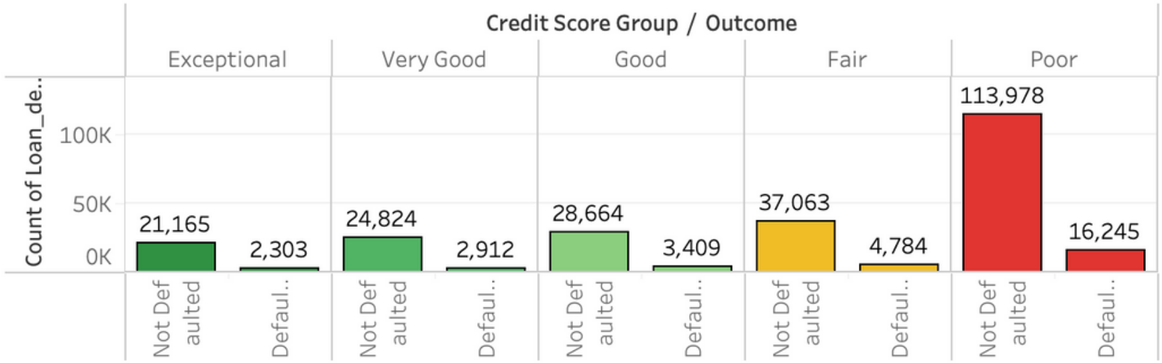
Groups based on Experian

- Exceptional: 800 - 850
- Very Good: 740 - 799
- Good: 670 - 739
- Fair: 580 - 669
- Poor: 300 - 579

Credit Score Analysis



Group Default Breakdown

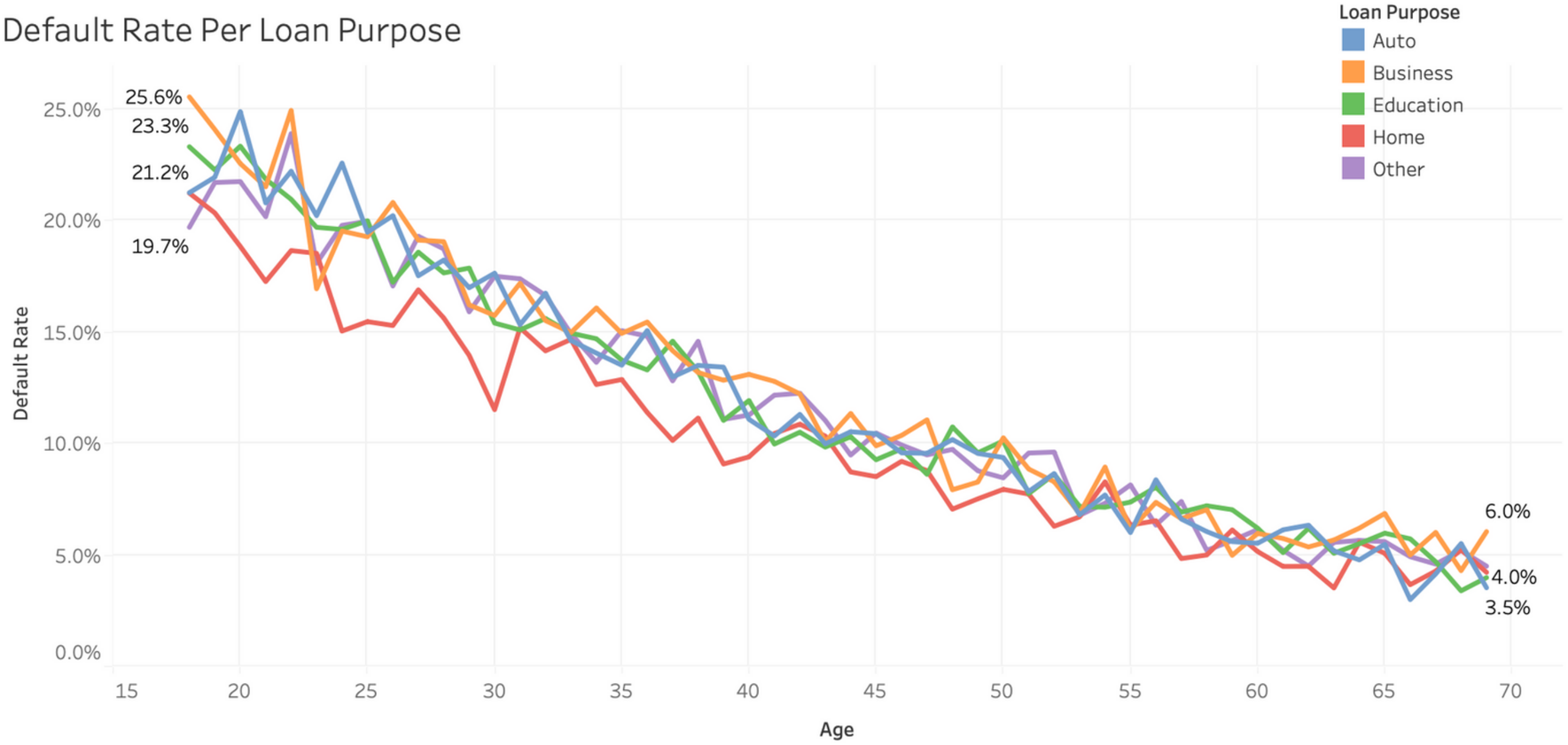


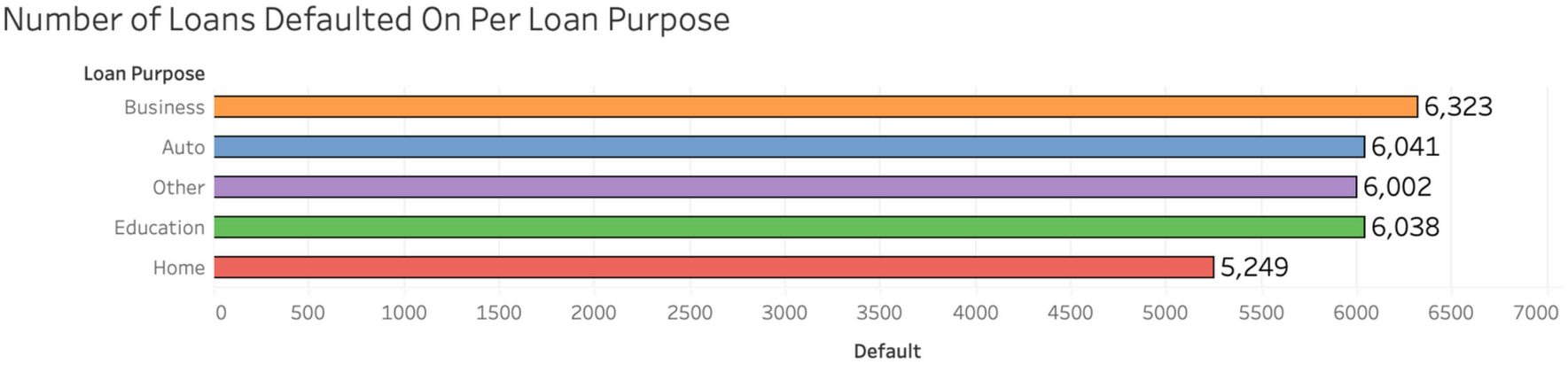
Group Default Rate

	Poor	Fair	Good	Very Good	Exceptional
Default Rate	12.5%	11.4%	10.6%	10.5%	9.8%

LOAN PURPOSE BREAKDOWN

- Overall downward trend as the age goes up
- Default Rate: Business> Auto > Education > Home



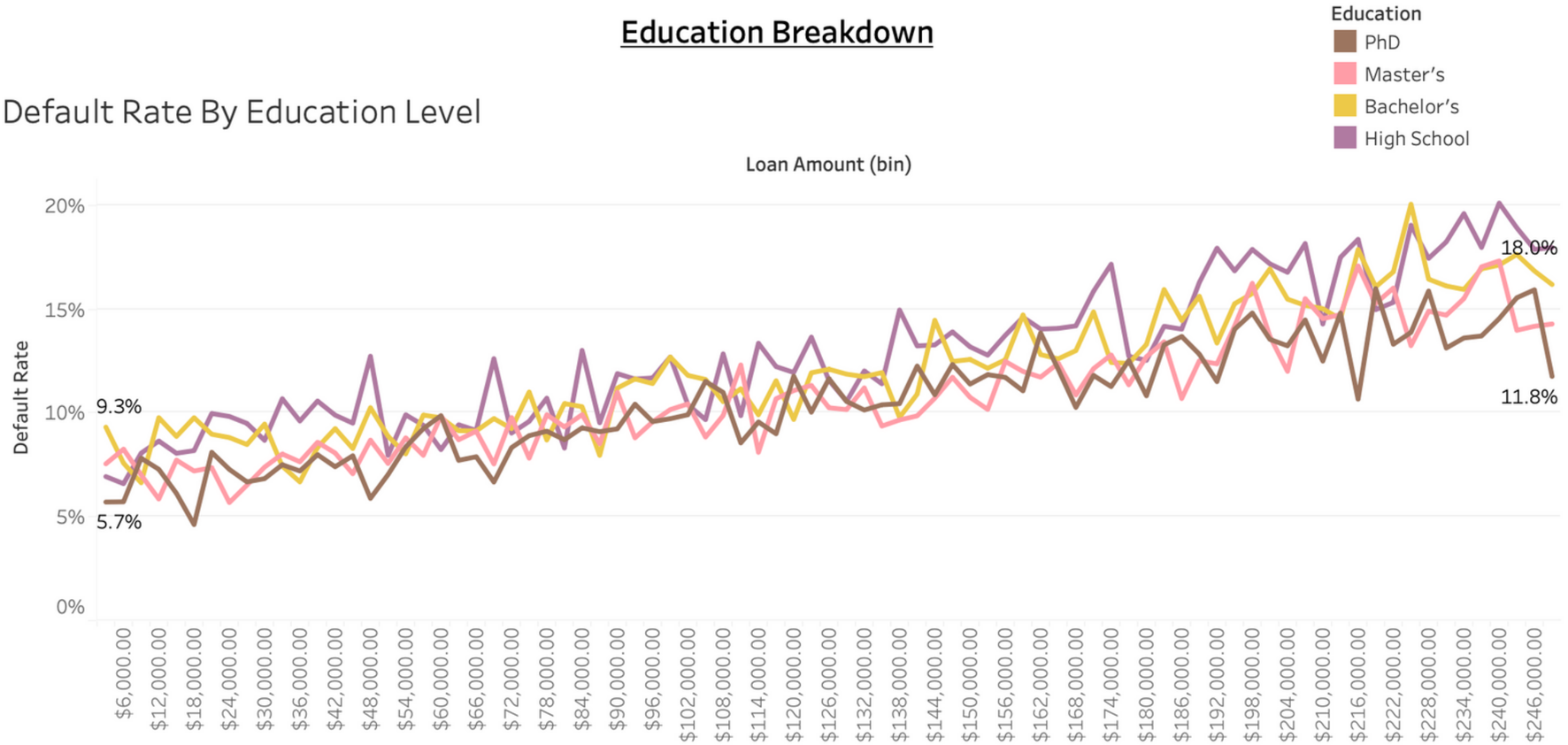


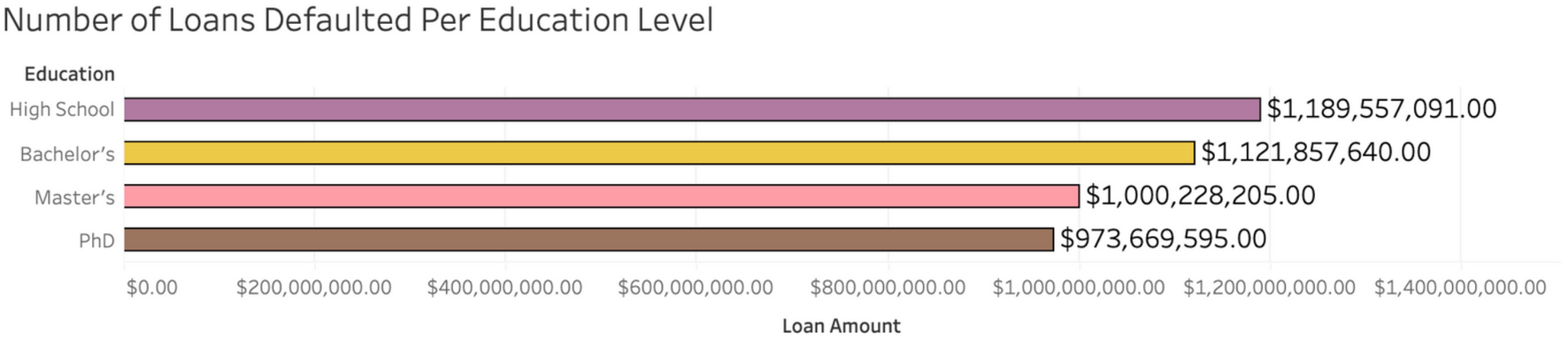
	Auto	Business	Education	Home	Other
Default Rate	11.9%	12.3%	11.8%	10.2%	11.8%

EDUCATION LEVEL BREAKDOWN

- Overall upward trend as the loan amount goes up
- Default Rate: High School> Bachelor’s > Master’s > PhD

Education Breakdown



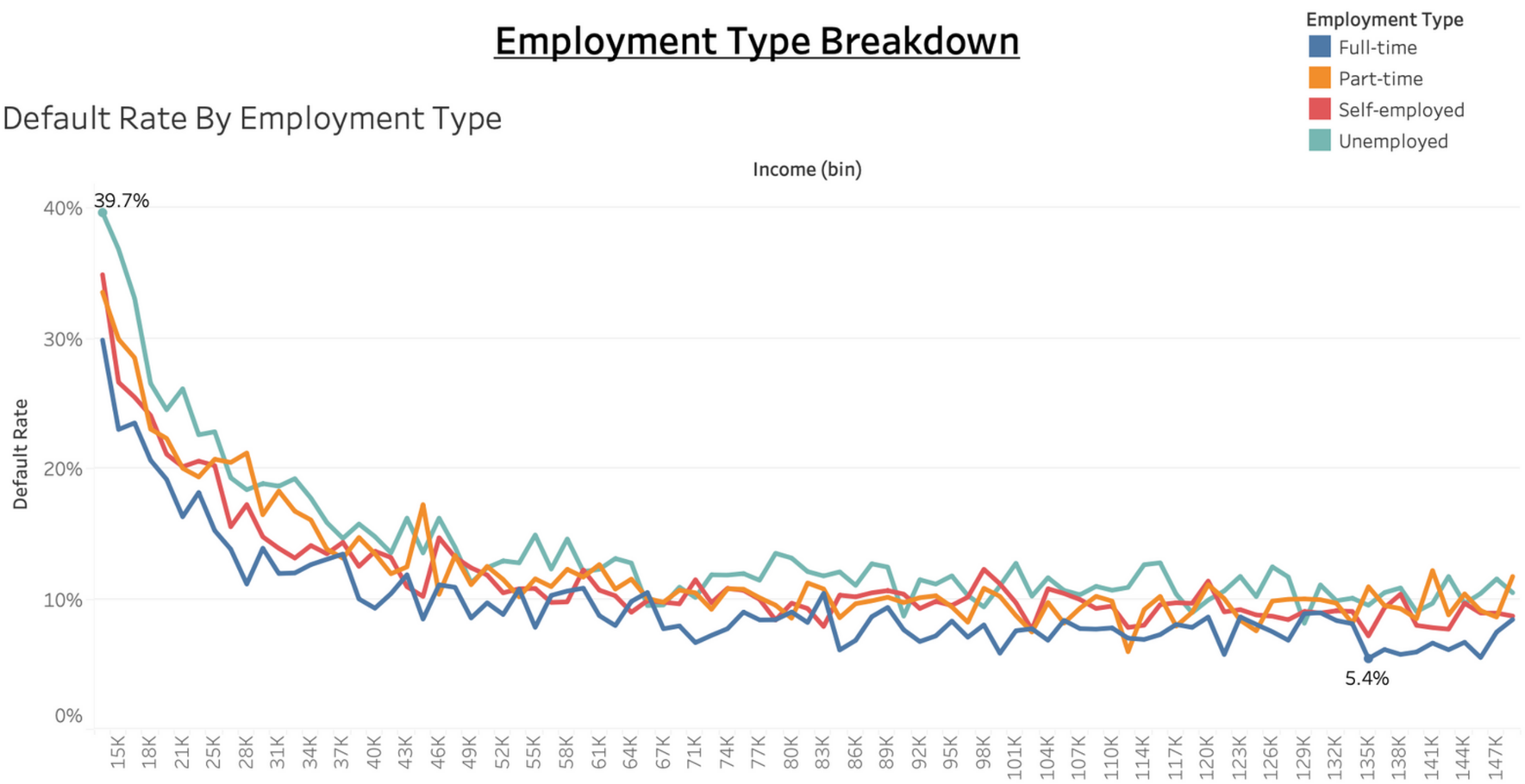


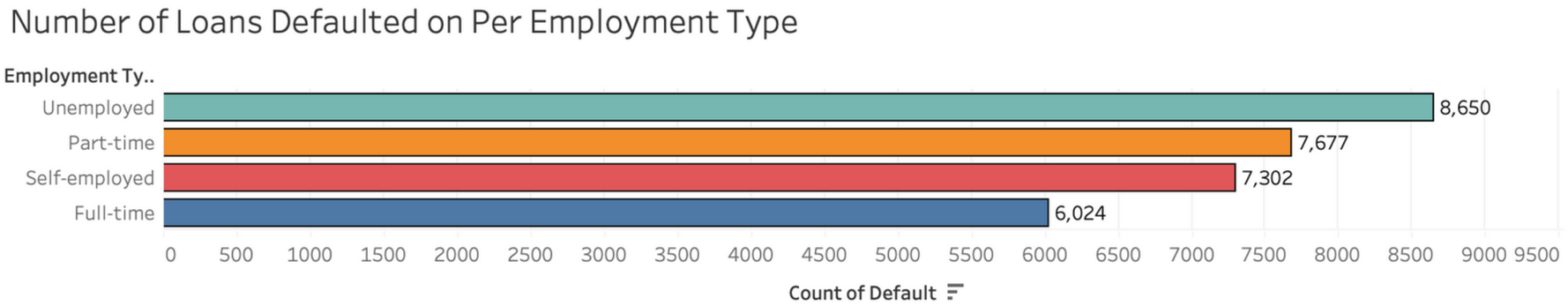
	PhD	Master’s	Bachelor’s	High School
Default Rate	10.6%	10.9%	12.1%	12.9%

EMPLOYMENT TYPE BREAKDOWN

- Overall downward trend as the salary amount goes up
- Default Rate: Unemployed> Part-time > Self-employed> Full-time

Employment Type Breakdown





	Full-time	Part-time	Self-employed	Unemployed
Default Rate	9.5%	12.0%	11.5%	13.6%

LOAN DEFAULT: LOGISTIC REGRESSION MODELLING

DATA PREPARATION

1. **Importing the dataset and checking for Nulls/Missing Values**
2. **Dropping Unnecessary Columns**
3. **Checking and Modifying Datatypes**
4. **Identifying Categorical Attributes and assigning Dummy Variables to the classes in each attribute**
5. **Checking for Correlation between Outcome and Predictors**
6. **Creating a separate Dataframe for predictors and an array for the Outcome**
7. **Splittling into Training and Testing Sets (60%-40%)**

OUTCOMES OF LOGISTIC REGRESSION

I1. In the Overall Logistic Model,
Age, Income, Loan Amount, Credit Score, Months Employed, Number of Credit Lines, Interest Rate and DTI Ratio have the highest significance in determining the Defaulting Probability
2. The model obtained using the training set has an accuracy of 88%
3. The model also has a prediction accuracy of 88% for the testing set.
4. Out of 102000 values, 90200 values were classified correctly by the prediction model
5. From the logit model, we deduce that:
Age, Income, Months Employed, Number of Credit Lines are inversely proportional with chances of Loan Defaulting
6. Interest Rate, DTI Ratio, Loan Amount , Loan Term vary proportionally with chances of Loan Defaulting

Confusion Matrix for Test Set

```
from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(Outcome_test, pred_Y)
print(confusion_matrix)
```

```
[[90129    87]
 [11759   164]]
```

Training Model Accuracy	Testing Model Accuracy
88%	88%

LOAN DEFAULT: DECISION TREE MODELLING

PREPARING THE DATA FOR MODELING

- Dropping Irrelevant Columns -'LoanID'
- Handling Missing Values - to ensure data completeness
- Converting categorical variables into numerical

```
# Convert categorical variables to numerical
loan_df = pd.get_dummies(loan_df, columns=['Education', 'EmploymentType', 'MaritalStatus', 'LoanPurpose', 'HasMortga
       'HasDependents', 'LoanPurpose', 'HasCoSigner'])
```

MODELING WITH DECISION TREE CLASSIFIER

- **‘train_test_split’** - 80% training, 20% testing
- **DecisionTreeClassifier** - max_depth=7
- ‘**clf.fit(X_train, y_train)**’ - finding patterns in the data

```
# Initialize the Decision Tree model
clf = DecisionTreeClassifier(max_depth = 7)

# Fit the model on the training data
clf.fit(X_train, y_train)
```

```
DecisionTreeClassifier(max_depth=7)
```

EVALUATING THE DECISION TREE

```
Accuracy: 0.8856993200409248
Confusion Matrix:
[[180111    413]
 [ 22936    817]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94    180524
           1       0.66      0.03      0.07     23753

    accuracy                           0.89    204277
   macro avg       0.78      0.52      0.50    204277
weighted avg       0.86      0.89      0.84    204277
```

- **overall accuracy of prediction on train data is 88.6%**
- **True Negatives - 180,111**

- **overall accuracy of prediction on test data is 88.5%**
- **True Negatives - 45,019**

```
Accuracy: 0.8852555316232622
Confusion Matrix:
[[45019   151]
 [ 5709   191]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      1.00      0.94     45170
           1       0.56      0.03      0.06      5900

    accuracy                           0.89     51070
   macro avg       0.72      0.51      0.50     51070
weighted avg       0.85      0.89      0.84     51070
```

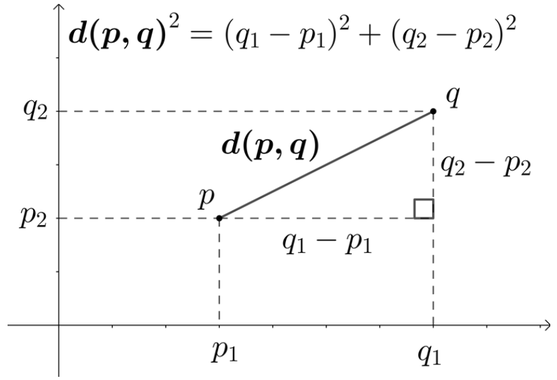
LOAN DEFAULT: K-NEAREST NEIGHBORS MODELLING

PREPARING THE DATA FOR MODELING

- Select numerical variables for distance calculation.
- Choose the best k-value for the model
- Using Euclidean distance for continuous variables

```
# Selecting numerical columns for KNN
numerical_cols = ['Age', 'Income', 'LoanAmount', 'CreditScore', 'MonthsEmployed',
                  'NumCreditLines', 'InterestRate', 'LoanTerm', 'DTIRatio']
numerical_data = data[numerical_cols]
```

```
# Applying K-Nearest Neighbors (KNN) with 5 neighbors
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(train_data_scaled, train_labels)
```



EVALUATING THE KNN MODEL

```
Training Set Evaluation:
Confusion Matrix:
 [[89257  1047]
 [ 9673  2161]]
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.99      0.94     90304
           1       0.67      0.18      0.29     11834

    accuracy                           0.90    102138
   macro avg       0.79      0.59      0.62    102138
weighted avg       0.88      0.90      0.87    102138
```

```
Testing Set Evaluation:
Confusion Matrix:
 [[132090   3300]
 [ 16328   1491]]
Classification Report:
              precision    recall  f1-score   support

           0       0.89      0.98      0.93    135390
           1       0.31      0.08      0.13     17819

    accuracy                           0.87    153209
   macro avg       0.60      0.53      0.53    153209
weighted avg       0.82      0.87      0.84    153209
```

These results suggest that the model fits the training data reasonably well, with a good balance between precision and recall for the non-default class (class 0).

However, for the default class (class 1), the recall is low, indicating that the **model struggles to identify default cases correctly even in the training set.**

The performance drop from training to testing suggests the model **may not be capturing the complexity** needed to distinguish between the classes effectively.

MODEL EVALUATION

MODEL ACCURACY

	Training Accuracy	Testing Accuracy
Logistic Regression	88%	88%
Decision Tree	88.6%	88.5%
k-NN	89.5%	87.1%

THANK YOU

FOR YOUR ATTENTION