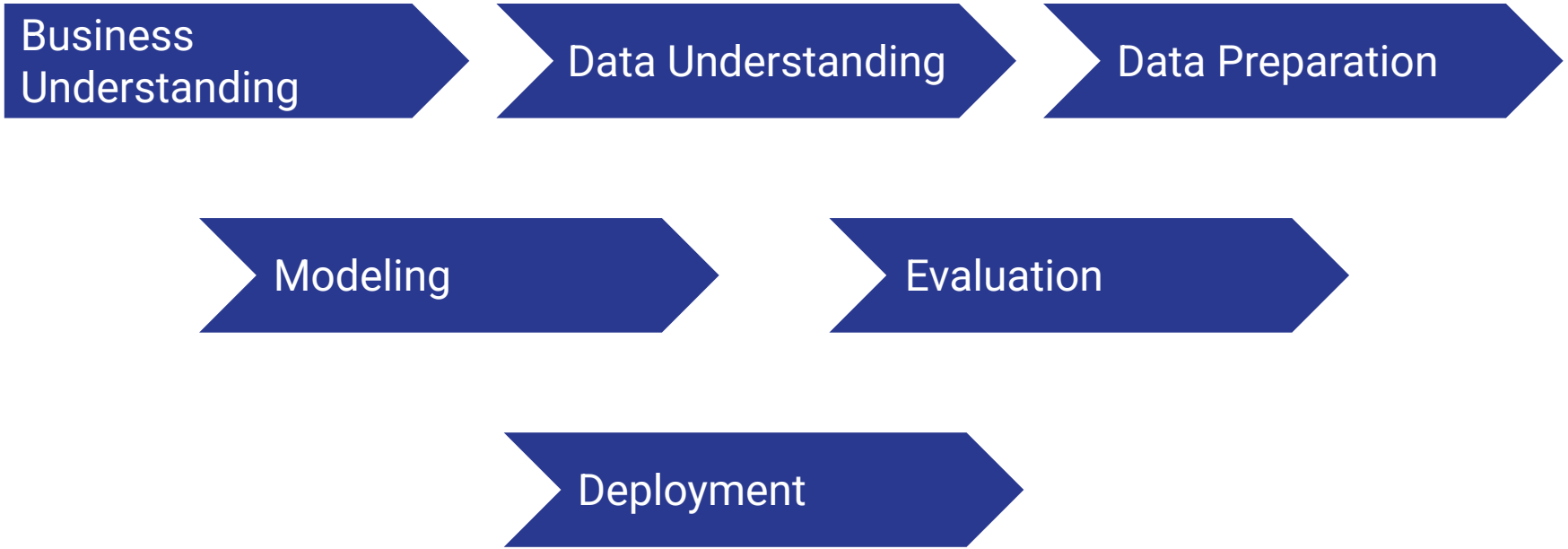
Lending Club Financial Data Case Study

Supervised Learning Application

Team Name: Data Whispers
Team Members:

- Kanyepi Ronald
- Zhao Runxun
- Musembi Evelyn
- Qian Tina

Agenda


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment

Business Understanding

Our goal is to create a supervised learning model that can forecast loan default rates for borrowers. It is essential to comprehend the elements that affect loan default in order to lower lending risks and make wise choices. We can assist lenders better manage credit risk by classifying loans as either defaulted or not by assessing borrower and loan information features.

Data Description

Loan Details

- **loan_amnt**: amount of the loan
- **term**: term of the loan
- **int_rate**: interest rate of the loan
- **installment**: monthly payment of the loan
- **purpose**: purpose of the loan
- **dti**:debt-to-income
- **delinq_2yrs**: number of delinquencies in the past 2 years
- **inq_last_6mths**: number of inquiries in the last 6 months
- **open_acc**: number of open credit lines in the borrower's credit file
- **pub_rec**:number of derogatory public records
- **revol_bal**: total credit revolving balance
- **revol_util**: amount of credit the borrower is using relative to their total credit limit
- **total_acc**: total number of credit lines in the borrower's credit file
- **last_pymnt_amnt**: last payment amount received

Borrower Details

- **grade**: the grade assigned by the lending institution based on the creditworthiness of the borrower
- **sub_grade**: more detailed grade based on the borrower's creditworthiness
- **emp_length**: length of employment of the borrower
- **home_ownership**: type of home ownership of the borrower
- **annual_inc**: annual income of the borrower
- **verification_status**: whether the income was verified by lending institution

Data Understanding - 142 columns in total

Causing Data Leakage

The potential to cause data leakage in the prediction of loan defaults. **(50)** e.g

hardship_amount,hardship_flag,hardship_reason

Irrelevant Features

Being unique identifiers or unrelated to loan performance. **(18)**

e.g. *member_id,policy_id*

Missing value

Features that have a high percentage of missing values.

(45) *features had more than 50% of their values missing.*

Target Variable Remapping

loan_status column's labels information:

- **Fully Paid**: The loan has been fully paid off by the borrower. **-> Non-default**
- **Charged Off**: The loan has not been fully repaid, and Lending Club has charged off the remaining balance as a loss. **-> Default**
- **Current**: The loan is currently being repaid on schedule. **-> Non-default**
- **Default**: The borrower has failed to make payments on the loan, and the loan is in default. **-> Default**
- **Late (31-120 days)**: The borrower has missed payments and is between 31 and 120 days late on their payment schedule. **-> Non-default**
- **In Grace Period**: The borrower is in a grace period and has missed a payment. **-> Non-default**
- **Late (16-30 days)**: The borrower has missed a payment and is between 16 and 30 days late on their payment schedule. **-> Non-default**
- **Does not meet the credit policy. Status: Fully Paid**: The loan has been fully paid off, but did not meet Lending Club's credit underwriting policy. **-> Non-default**
- **Does not meet the credit policy. Status: Charged Off**: The loan did not meet Lending Club's credit underwriting policy and has been charged off as a loss. **-> Default**
- **Issued**: The loan has been issued but has not yet been funded by disbursed. **-> Non-default**

Binary Classification

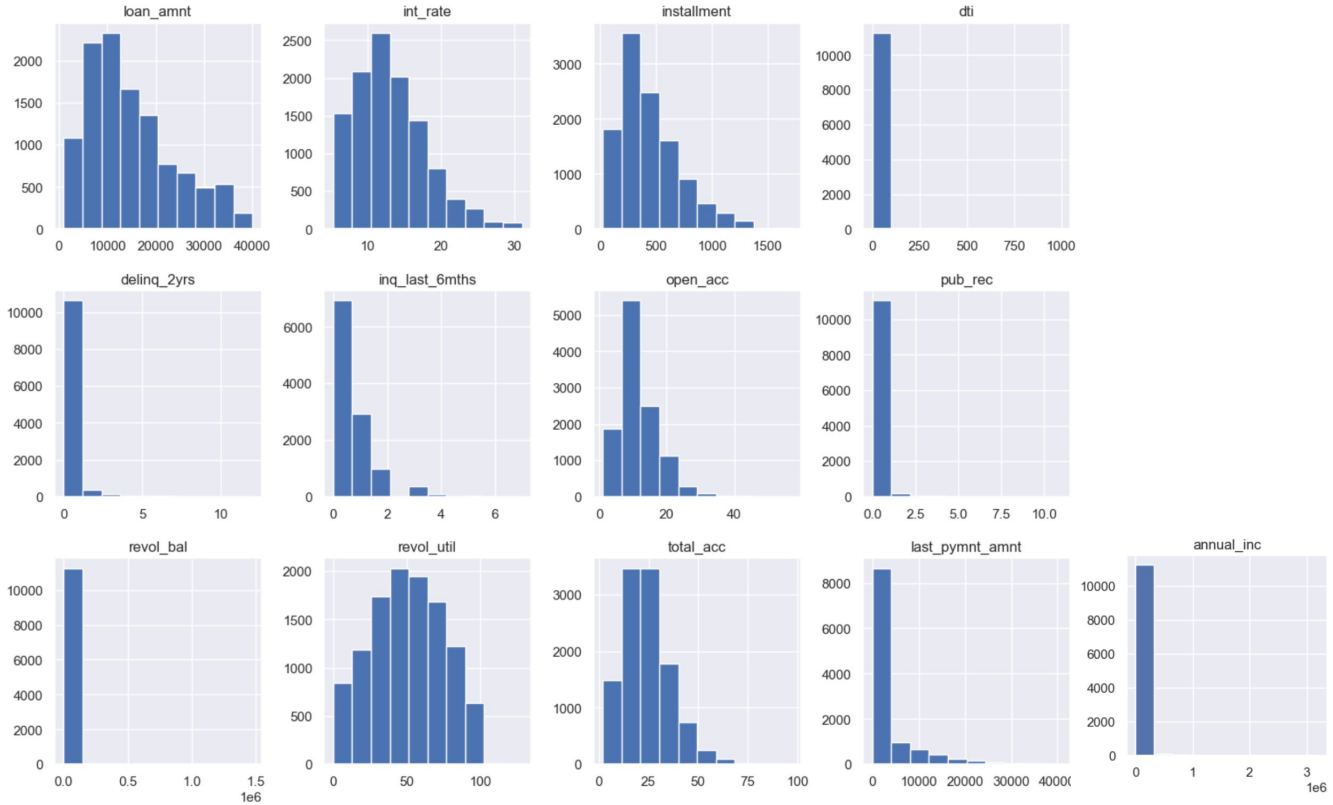
loan_status_	count
Non-default	9943
Default	1361

loan_status	count
Charged Off	1356
Current	4380
Does not meet the credit policy. Status:Charged Off	5
Does not meet the credit policy. Status:Fully Paid	5
Fully Paid	5377
In Grace Period	42
Late (16-30 days)	26
Late (31-120 days)	113

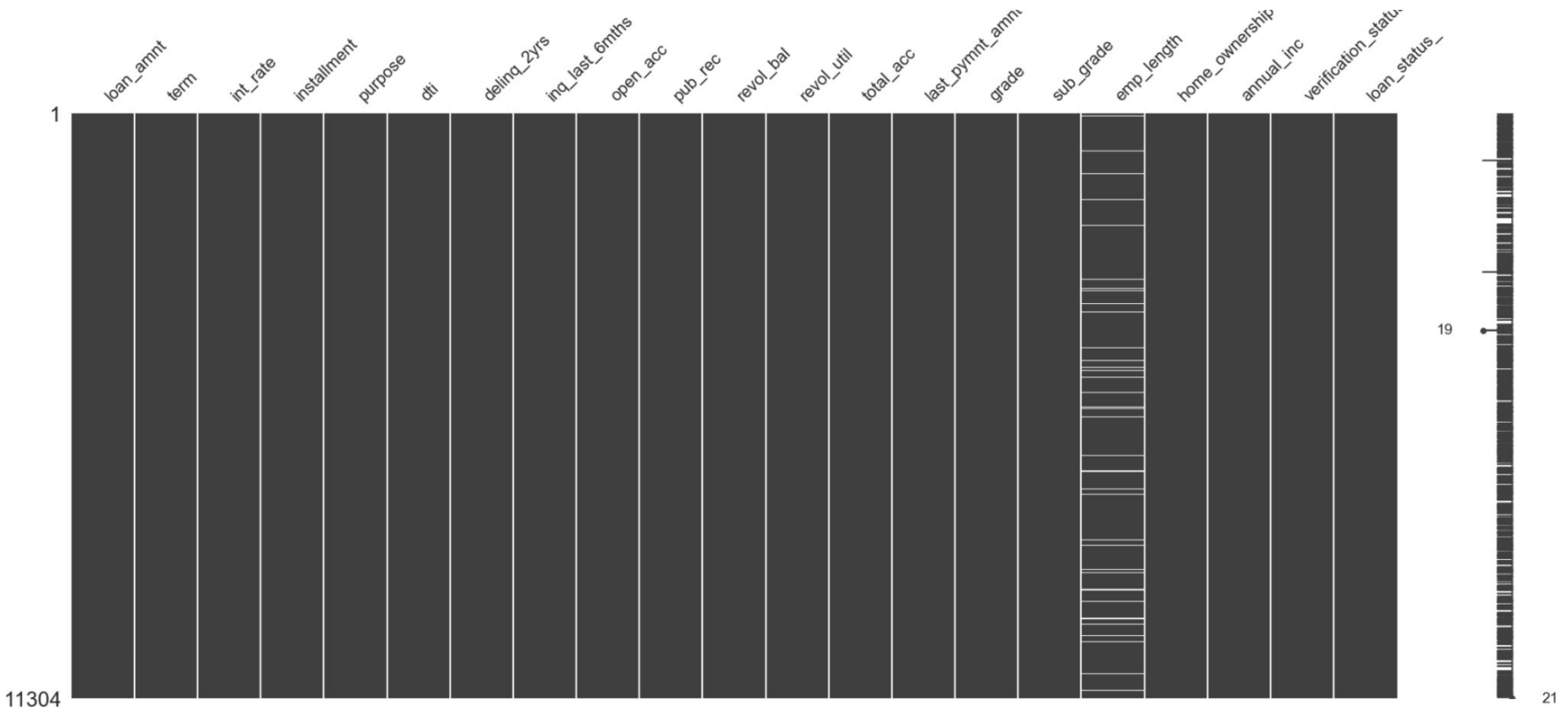
Summary Statistics

	loan_amnt	int_rate	installment	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util
count	11304.000000	11304.000000	11304.000000	11302.000000	11304.000000	11304.000000	11304.000000	11304.000000	1.130400e+04	11299.000000
mean	14959.509908	13.059967	442.754882	18.937142	0.302105	0.569798	11.584660	0.191791	1.670841e+04	50.518515
std	9209.509310	4.840431	266.247443	16.763837	0.839258	0.870797	5.683574	0.530766	2.649104e+04	24.747123
min	1000.000000	5.310000	23.010000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000e+00	0.000000
25%	8000.000000	9.440000	249.080000	11.790000	0.000000	0.000000	8.000000	0.000000	5.934000e+03	32.000000
50%	12800.000000	12.620000	376.210000	17.800000	0.000000	0.000000	10.000000	0.000000	1.134400e+04	50.500000
75%	20000.000000	15.880000	586.285000	24.470000	0.000000	1.000000	14.000000	0.000000	2.008300e+04	69.600000
max	40000.000000	30.990000	1714.540000	999.000000	12.000000	7.000000	57.000000	11.000000	1.470945e+06	128.600000

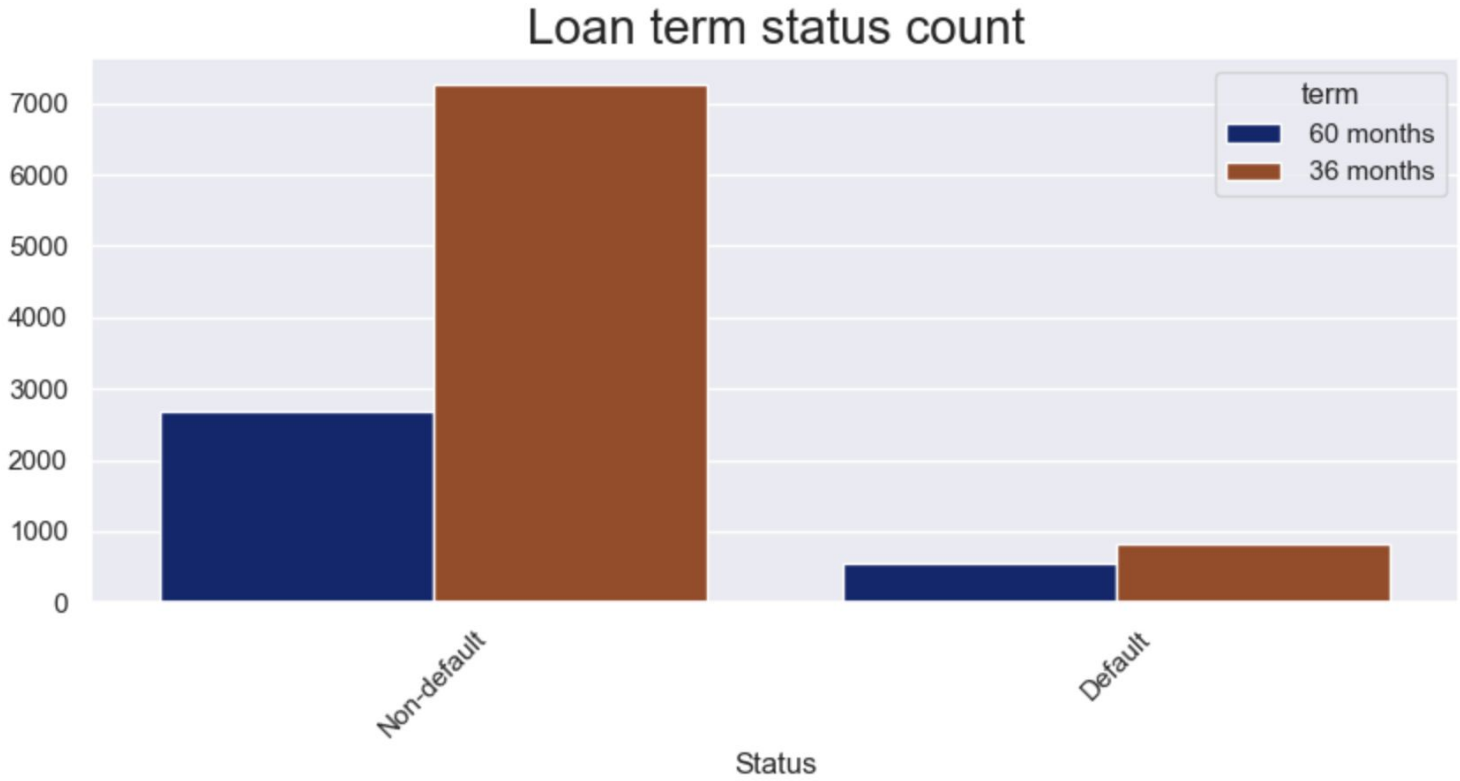
Distribution of features

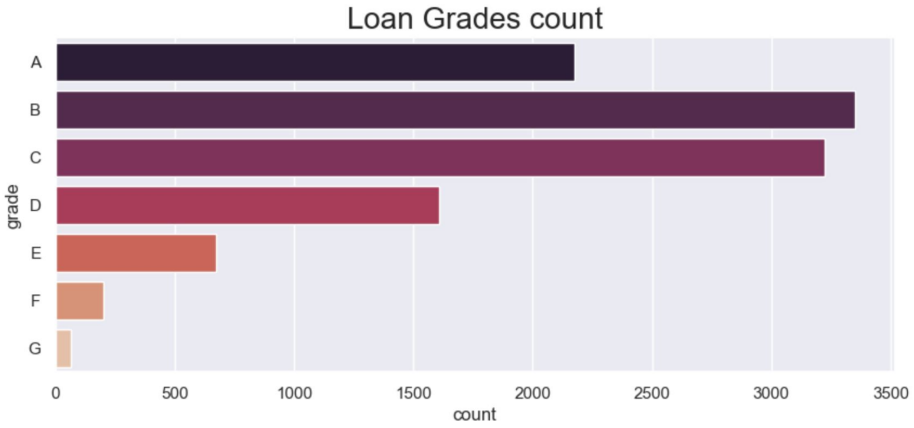


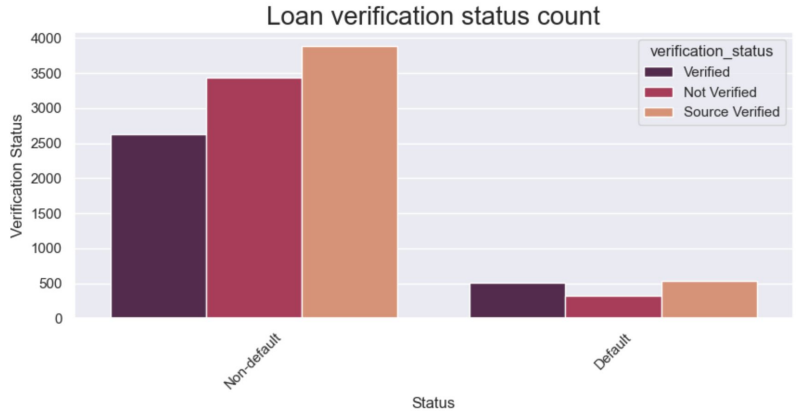
Missing value

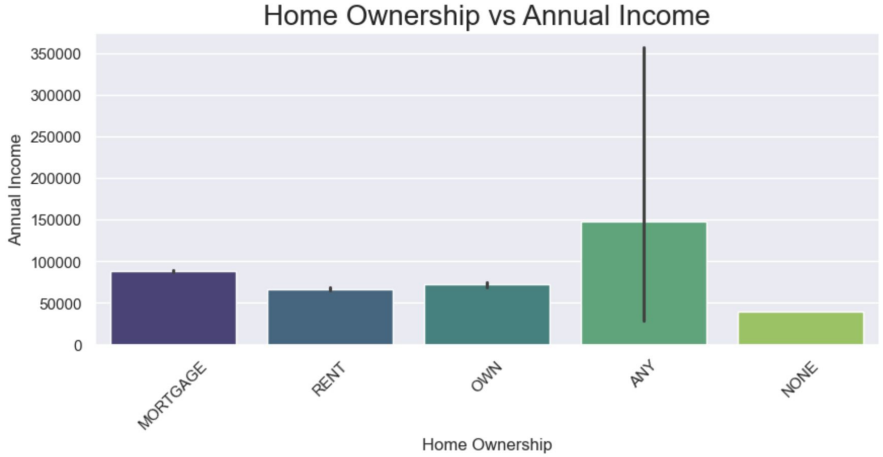


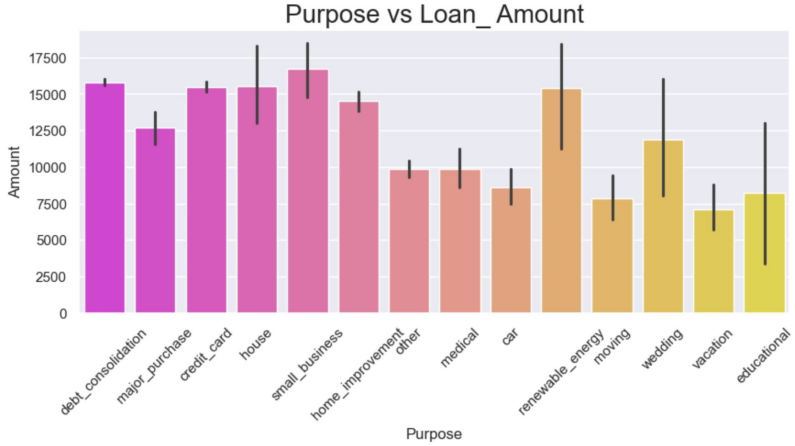
Exploratory Data Analysis

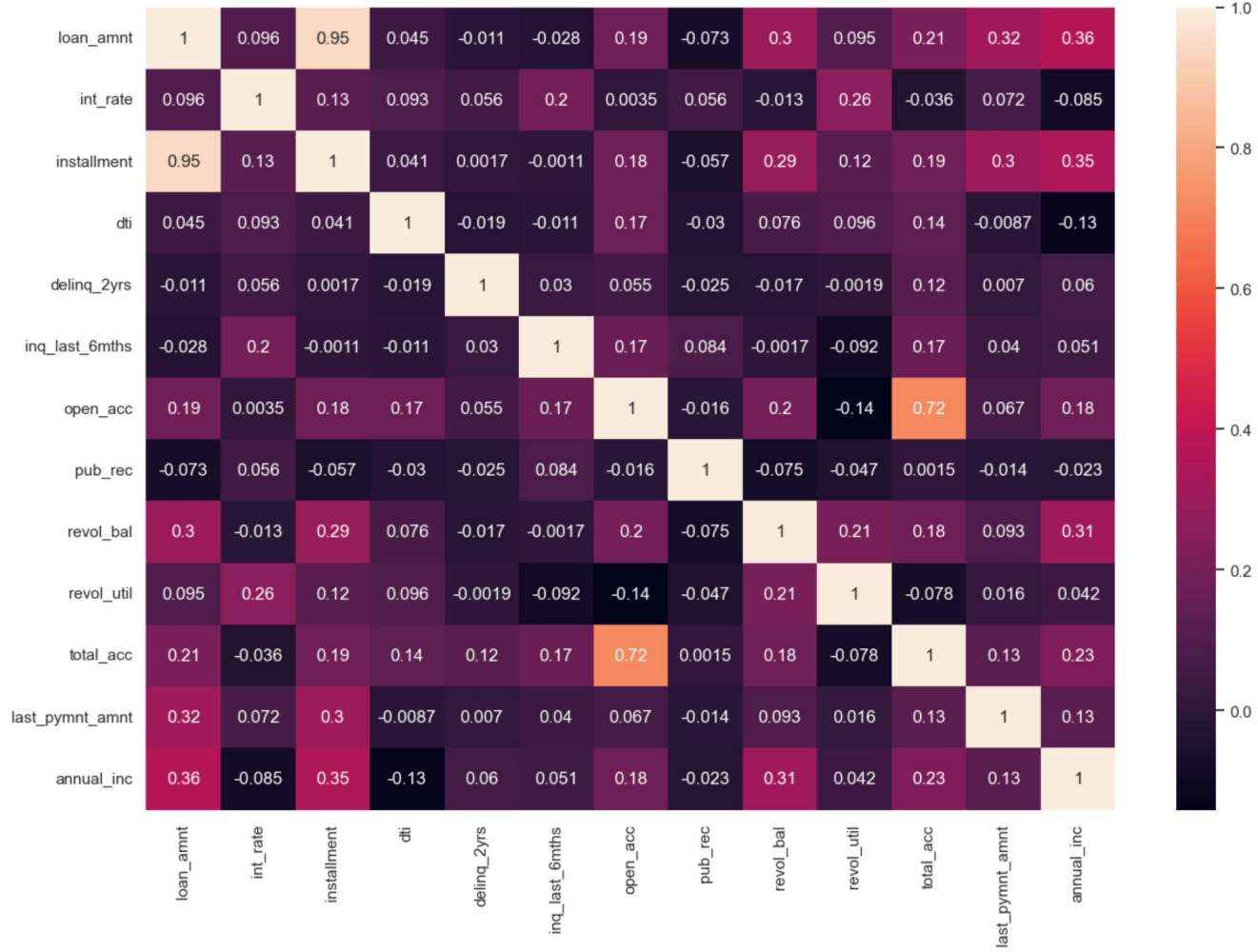



Loan Grades count
grade
A
B
C
D
E
F
G
0
500
1000
1500
2000
2500
3000
3500
count


Loan verification status count
verification_status
Verified
Not Verified
Source Verified
Verification Status
4000
3500
3000
2500
2000
1500
1000
500
0
Non-default
Default
Status


Home Ownership vs Annual Income
Annual Income
350000
300000
250000
200000
150000
100000
50000
0
MORTGAGE
RENT
OWN
ANY
NONE
Home Ownership


Purpose vs Loan_ Amount
Amount
17500
15000
12500
10000
7500
5000
2500
0
debt_consolidation
major_purchase
credit_card
house
small_business
home_improvement
other
medical
car
renewable_energy
moving
wedding
vacation
educational
Purpose


loan_amnt 1 0.096 0.95 0.045 -0.011 -0.028 0.19 -0.073 0.3 0.095 0.21 0.32 0.36
int_rate 0.096 1 0.13 0.093 0.056 0.2 0.0035 0.056 -0.013 0.26 -0.036 0.072 -0.085
installment 0.95 0.13 1 0.041 0.0017 -0.0011 0.18 -0.057 0.29 0.12 0.19 0.3 0.35
dti 0.045 0.093 0.041 1 -0.019 -0.011 0.17 -0.03 0.076 0.096 0.14 -0.0087 -0.13
delinq_2yrs -0.011 0.056 0.0017 -0.019 1 0.03 0.055 -0.025 -0.017 -0.0019 0.12 0.007 0.06
inq_last_6mths -0.028 0.2 -0.0011 -0.011 0.03 1 0.17 0.084 -0.0017 -0.092 0.17 0.04 0.051
open_acc 0.19 0.0035 0.18 0.17 0.055 0.17 1 -0.016 0.2 -0.14 0.72 0.067 0.18
pub_rec -0.073 0.056 -0.057 -0.03 -0.025 0.084 -0.016 1 -0.075 -0.047 0.0015 -0.014 -0.023
revol_bal 0.3 -0.013 0.29 0.076 -0.017 -0.0017 0.2 -0.075 1 0.21 0.18 0.093 0.31
revol_util 0.095 0.26 0.12 0.096 -0.0019 -0.092 -0.14 -0.047 0.21 1 -0.078 0.016 0.042
total_acc 0.21 -0.036 0.19 0.14 0.12 0.17 0.72 0.0015 0.18 -0.078 1 0.13 0.23
last_pymnt_amnt 0.32 0.072 0.3 -0.0087 0.007 0.04 0.067 -0.014 0.093 0.016 0.13 1 0.13
annual_inc 0.36 -0.085 0.35 -0.13 0.06 0.051 0.18 -0.023 0.31 0.042 0.23 0.13 1
loan_amnt
int_rate
installment
dti
delinq_2yrs
inq_last_6mths
open_acc
pub_rec
revol_bal
revol_util
total_acc
last_pymnt_amnt
annual_inc
1.0
0.8
0.6
0.4
0.2
0.0

Data Preparation

Data Cleaning & Wrangling

Handling Features:

- Removed columns with over 50% missing data, irrelevant features, and those causing data leakage.
- Applied domain knowledge to select relevant features for further processing.

Imputing Missing Values:

- Mean for numeric variables
- Most frequent for categorical variables

Target Variable Mapping

Created a new target variable `loan_status_` and mapped it as follows

- Non-default: 0, Default: 1.

One-Hot Encoding : for Nominal categorical variables e.g purpose

Label Encoder : for Ordinal Categorical variables e.g grade

Train-Test Split & Dataset Balancing

Splitting: Dataset split into 80% training and 20% testing using stratification.

Balancing: Applied RandomUnderSampler to balance the training dataset.

Modeling

Decision Tree

Model:	Decision Tree
Non-nested CV F1 Score:	0.73479116
Optimal Parameter:	{'criterion': 'gini', 'max_depth': 7, 'min_samples_split': 7}
Optimal Estimator:	DecisionTreeClassifier(max_depth=7, min_samples_split=7, random_state=42)
Nested CV F1 Score:	0.72275046
Nested CV F1 Score Std:	0.01591427

KNN

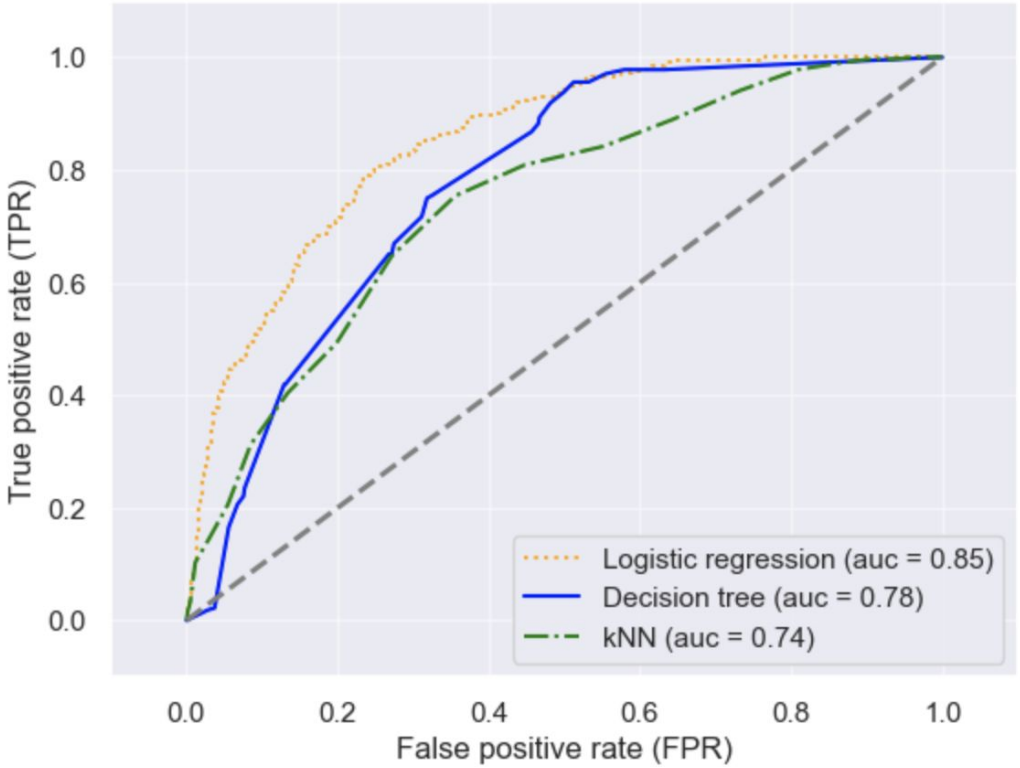
```
Model:                          KNN
Non-nested CV F1-Score:         0.6552
Optimal Parameter:              {'knn__n_neighbors': 21, 'knn__weights': 'uniform'}
Optimal Estimator:              Pipeline(steps=[('sc', StandardScaler()),
                ('knn', KNeighborsClassifier(n_neighbors=21))])
Nested CV F1-Score:             0.6539
Nested CV F1-Score Std:         0.0112
```

Logistic Regression

```
Nested CV F-1 Score: 0.7658626133021758  +/-  0.014796435695434968
Model:                          Logistic Regression
Non-nested CV F-1 Score:        0.76755746
Optimal Parameter:              {'fit_intercept': True, 'tol': 0.001}
Optimal Estimator:              LogisticRegression(multi_class='ovr', penalty='l1', random_state=42,
                   solver='liblinear', tol=0.001)
Nested CV F1 Score:             0.76586261
Nested CV F1 Score Std:         0.01479644
```

Model Evaluation

ROC Curves


True positive rate (TPR)
False positive rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Logistic regression (auc = 0.85)
Decision tree (auc = 0.78)
kNN (auc = 0.74)

Developing a Business Case for Expected Improvement:

- **Default Reduction:** Predict fewer defaults, saving financial losses.
- **Revenue Growth:** Increase approved loans for low-risk borrowers, boosting loan volume.
- **Cost Savings:** Reduce manual assessments and lower bad debt.
- **ROI:** Compare cost savings and additional revenue to model development and deployment costs.

Challenges in Projecting ROI

Challenges in Projecting ROI:

- **Uncertainty in default rates**
- **Complex cost attribution**

Alternatives if ROI is Difficult to Measure:

- **Risk-adjusted returns**
- **Qualitative benefits**

Deployment Potential

Loan Approval System:

- Automatically evaluate loan applications based on borrower details and loan features.
- Logistic regression provides robust predictions, ideal for large-scale deployment.
- Decision trees offer interpretability for decision-making support.

Credit Risk Management:

- Logistic regression helps assess the likelihood of loan default, improving risk mitigation.
- Suitable for financial institutions managing large customer bases.

Customer Segmentation for Lending Products:

- Decision trees can segment borrowers into risk categories for tailored loan offerings.
- Helps institutions adjust interest rates or loan terms based on borrower risk.

Deployment Issues to Consider:

- System Integration
- Data Privacy and Security
- Model Monitoring

Ethical Considerations:

- Bias in Predictions
- Transparency

Risks and Mitigation:

- Model Performance Drift
- Over-reliance on Automation
- Regulatory Compliance

Thank you!