Rakamin Academy

id/x partners

Credit Risk Analysis

VIX Data Scientist ID/X Partners

Presented by
Giselle Halim

LinkedIn GitHub

id/x partners

Rakamin Academy

Case Study

About ID/X Partners

Founded in 2002 by seasoned ex-bankers and management consultants, ID/X Partners (PT IDX Consulting) boasts extensive experience across various industry sectors. This expertise encompasses credit cycle and process management, scoring development, and performance management. Their proven track record spans companies in Asia and Australia, across diverse fields including financial services, telecommunications, manufacturing, and retail.

ID/X Partners specializes in data analytics and decisioning (DAD) solutions. These solutions are seamlessly integrated with risk management and marketing disciplines to empower clients to optimize portfolio profitability and business processes. This comprehensive approach, coupled with their diverse technology solutions, positions ID/X Partners as a one-stop shop for clients seeking data-driven success.

Case Study

A lending company, sought to enhance their risk assessment capabilities by developing a predictive model capable of assessing creditworthiness. The model would leverage a dataset comprising accepted and rejected loan applications to identify patterns and correlations indicative of credit risk. **The project aimed to provide a comprehensive solution encompassing model development, evaluation, and visual representation of findings, facilitating informed decision-making for the lending company.**

The timely development of this predictive model was crucial for the lending company to mitigate financial losses arising from loan defaults. By accurately identifying high-risk borrowers, the company could implement effective risk management strategies, optimize lending decisions, and improve overall profitability. Additionally, through Exploratory Data Analysis (EDA), **we can uncover patterns and trends within the data that reveal deeper insights into the factors influencing creditworthiness.**

Tools & Libraries

Tools:

- Python
- Google Colab

Libraries:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Scikit-learn

id/x partners

Rakamin Academy

Data Cleaning

Data Cleaning

```
#Deleting unused columns
dropped = ['id', 'member_id', 'Unnamed: 0', 'funded_amnt', 'funded_amnt_inv', 'url', 'desc', 'title', 'zip_code',
           'next_pymnt_d', 'last_pymnt_d', 'mths_since_last_delinq', 'mths_since_last_record', 'recoveries',
           'collection_recovery_fee', 'last_pymnt_d', 'next_pymnt_d', 'last_credit_pull_d',
           'mths_since_last_major_derog', 'total_pymnt', 'total_pymnt_inv', 'total_rec_prncp', 'total_rec_int',
           'out_prncp_inv', 'total_rev_hi_lim']

df.drop(columns = dropped, axis=1, inplace=True)

#Deleting all columns with null value above 90%
df = df.loc[:, df.isnull().mean() < .9]
```

Unused columns should be removed due to irrelevance, privacy concerns, or redundancy. Columns like id, member_id, url, desc, and title contain unique identifiers or descriptive text that do not contribute to predicting loan outcomes. Similarly, zip code offers limited predictive power in this context. Date-related columns like next_pymnt_d, last_pymnt_d, and last_credit_pull_d can be replaced with issue_d. Redundant columns such as funded_amnt, funded_amnt_inv, total_pymnt, and total_pymnt_inv can be substituted with loan_amnt. **Columns with over 90% missing data should be deleted to prevent hindrance in the data modeling process.**

Data Cleaning

loan_status	count
Current	224226
Fully Paid	184739
Charged Off	42475
Late (31-120 days)	6900
In Grace Period	3146
Does not meet the credit policy. Status:Fully Paid	1988
Late (16-30 days)	1218
Default	832
Does not meet the credit policy. Status:Charged Off	761

```
good_loan = ['Fully Paid', 'Does not meet the credit policy. Status:Fully Paid', 'In Grace Period', 'Late (16-30 days)']
df['loan_stats'] = np.where(df['loan_status'].isin(good_loan), 0, 1)

df = df.loc[~df['loan_status'].isin(['Current'])].reset_index(drop=True)
df.info()
```

To streamline the model's learning process and enhance prediction accuracy, the loan_status variable was categorized into two distinct classes. **0: Good (loans with timely payments or specific exemptions)** and **1: Bad (loans with overdue payments exceeding 30 days).** By reducing the number of categories, the model could focus on identifying the most significant factors influencing credit worthiness, minimizing the risk of overfitting and improving overall predictive performance. **The 'Current' loan status category is excluded from the analysis as these loans are ongoing and cannot be predicted.** The 'Late (16-30 days)' category is treated as a good loan since loans are typically not classified as bad until they surpass 30 days overdue.

Data Cleaning

```
df['policy_code'].unique()
```

```
array([1])
```

```
#Policy code only has 1 unique value
df.drop('policy_code', axis=1, inplace=True)
```

```
#Application type only has 1 unique value
df['application_type'].unique()
```

```
array(['INDIVIDUAL'], dtype=object)
```

```
df.drop('application_type', axis=1, inplace=True)
```

```
pymnt_plan
n    242052
y         7
Name: count, dtype: int64
```

```
#Dominated by a single value
df.drop('pymnt_plan', axis=1, inplace=True)
```

To enhance model performance and reduce dimensionality, columns containing one unique values were eliminated. Additionally, the pymnt_plan column, characterized by a significant preponderance of a single value, was removed due to its limited predictive power. This data cleansing process streamlined the dataset, focusing on features with meaningful variation that could contribute to accurate credit risk prediction.

Data Cleaning

```
df['emp_length'].unique()
```

```
array(['10+ years', '< 1 year', '1 year', '3 years', '8 years', '9 years',
       '4 years', '5 years', '6 years', '2 years', '7 years', nan],
      dtype=object)
```

```
#Cleaning emp_length values
df['emp_length'] = df['emp_length'].str.replace('\+ years', '')
df['emp_length'] = df['emp_length'].str.replace('< 1 year', str(0))
df['emp_length'] = df['emp_length'].str.replace(' years', '')
df['emp_length'] = df['emp_length'].str.replace(' year', '')
df['emp_length'] = df['emp_length'].str.replace('+', '')
df['emp_length'] = df['emp_length'].astype(float)
```

```
#Converting emp_title values to uppercase
df['emp_title'] = df['emp_title'].str.upper()
```

```
#Cleaning term values
df['term'] = df['term'].str.replace(' months', '')
df['term'] = df['term'].astype(float)
```

To ensure the accuracy and reliability of the model's predictions, the emp_title, emp_length, and term columns were tidied. This involved converting year and month strings into purely numeric formats, eliminating any extraneous characters or formatting that could potentially hinder the model's training process. By standardizing these variables, the model could effectively process and analyze the data, leading to more accurate and robust predictions.

Data Cleaning

```
df['issue_d'] = pd.to_datetime(df['issue_d'], format='%b-%y')
df['issue_d'].head()
```

```
0   2011-12-01
1   2011-12-01
2   2011-12-01
3   2011-12-01
4   2011-12-01
Name: issue_d, dtype: datetime64[ns]
```

```
df['mths_since_issue_d'] = round(pd.to_numeric((pd.to_datetime('2016-12-31') - df['issue_d']) / np.timedelta64(1, 'M')))
df['mths_since_issue_d'].head()
```

```
df['earliest_cr_line'] = pd.to_datetime(df['earliest_cr_line'], format='%b-%y')
df['earliest_cr_line'].head()
```

```
0   1985-01-01
1   1999-04-01
2   2001-11-01
3   1996-02-01
4   1996-01-01
Name: earliest_cr_line, dtype: datetime64[ns]
```

```
df['mths_since_earliest_cr_line'] = round(pd.to_numeric((pd.to_datetime('2016-12-31') - df['earliest_cr_line']) / np.timedelta64(1, 'M')))
df['mths_since_earliest_cr_line'].head()
```

```
# Extract the year from 'issue_d'
df['issue_year'] = pd.to_datetime(df['issue_d']).dt.year
```

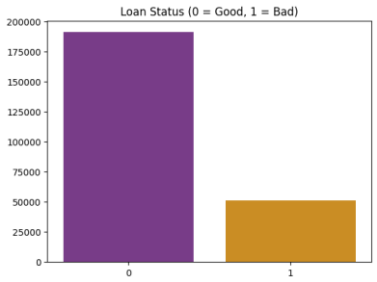
New features are created for analysis, such as calculating the months since issue_d and earliest_cr_line. A column containing the year of issue is also added for exploratory data analysis purposes.

id/x partners

Rakamin Academy

Exploratory Data Analysis

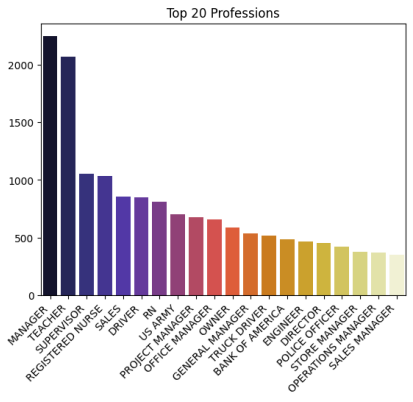
Exploratory Data Analysis



loan_stats	loan_amnt	term	int_rate
0	13233.481169	40.452706	13.323055
1	14596.853908	44.550306	15.973236

- **The data shows a skewed distribution where significantly more borrowers have good credit than bad credit**, which may lead to biased model predictions if not addressed through techniques like resampling.
- There are 191,091 good loans compared to 50,968 bad loans. **The percentage of bad loans to the total number of loans is approximately 21.06%. This is quite high and a potential red flag under most normal conditions.** It may indicate significant financial stress or a risky lending environment.
- The average loan amount is around $13-14k with an interest rate of 13-15%. **Bad loans tend to have a higher average loan amount than good loans, possibly indicating overborrowing by high-risk borrowers.**

Exploratory Data Analysis



Top 20 'emp_title' for loan_stats 0:

emp_title	
MANAGER	1596
TEACHER	1591
REGISTERED NURSE	782
SUPERVISOR	764
RN	626
SALES	615
DRIVER	564
US ARMY	557
PROJECT MANAGER	547
OFFICE MANAGER	511
BANK OF AMERICA	387
GENERAL MANAGER	382
ENGINEER	382
OWNER	377
DIRECTOR	365
POLICE OFFICER	350
TRUCK DRIVER	346
VICE PRESIDENT	302
OPERATIONS MANAGER	293
STORE MANAGER	273

Name: count, dtype: int64

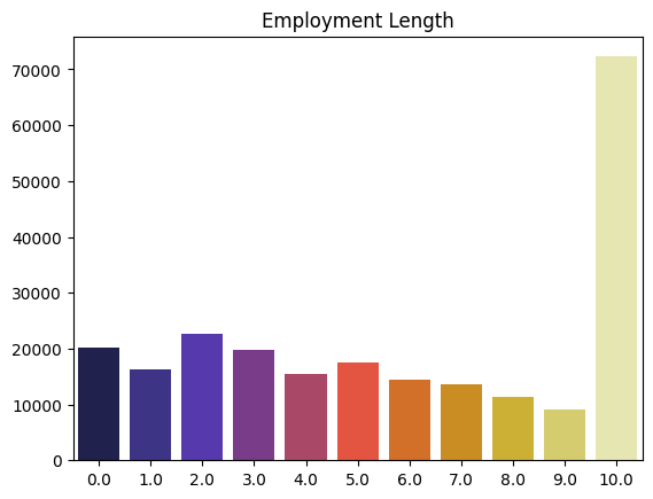
Top 20 'emp_title' for loan_stats 1:

emp_title	
MANAGER	647
TEACHER	479
SUPERVISOR	293
DRIVER	283
REGISTERED NURSE	253
SALES	243
OWNER	213
RN	186
TRUCK DRIVER	171
GENERAL MANAGER	156
OFFICE MANAGER	145
US ARMY	145
PROJECT MANAGER	130
STORE MANAGER	108
ASSISTANT MANAGER	98
NURSE	98
BANK OF AMERICA	98
SALES MANAGER	94
ENGINEER	88
DIRECTOR	87

Name: count, dtype: int64

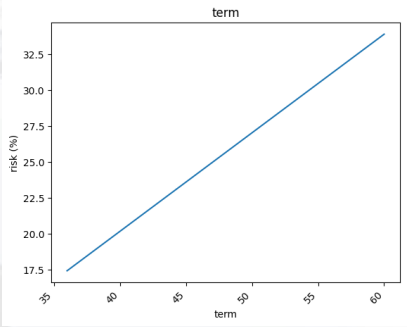
Managers and teachers are among the most frequent borrowers and are more likely to default on loans. This may stem from job-related financial strain, such as underpaid work or responsibilities requiring personal financial contributions (in the case of teachers), or from overextension of credit and debt accumulation due to higher living expenses or lifestyle demands for managers.

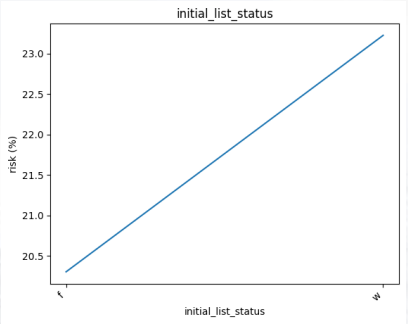
Exploratory Data Analysis

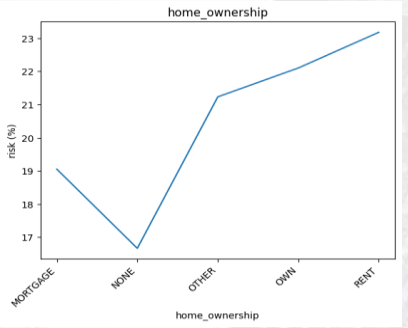


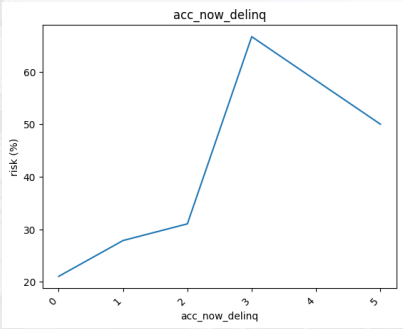
The average borrower has been employed for 10 or more years, but this does not necessarily indicate lower credit risk. Despite long work histories, other factors may influence default risk.

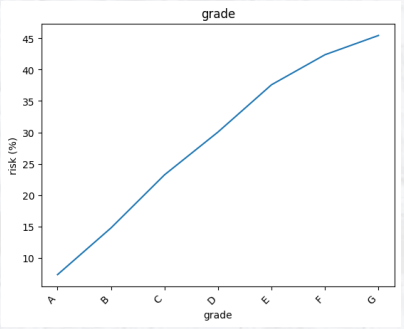
Exploratory Data Analysis

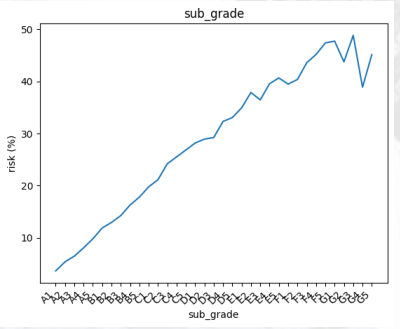




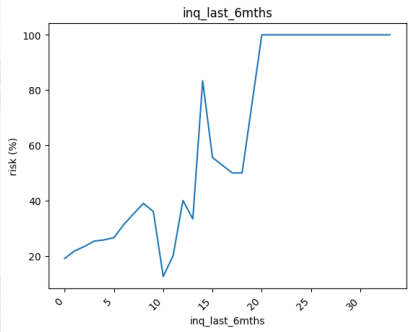


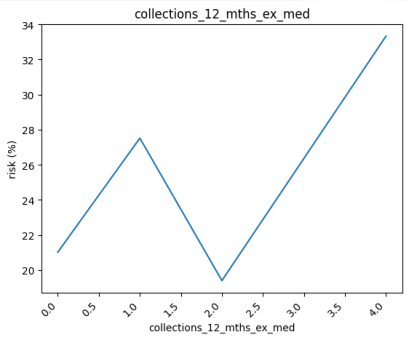


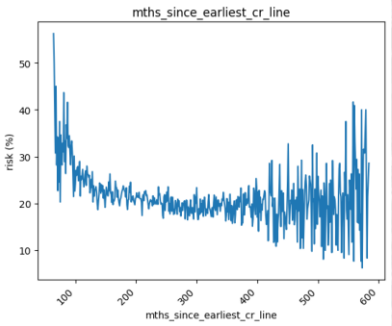


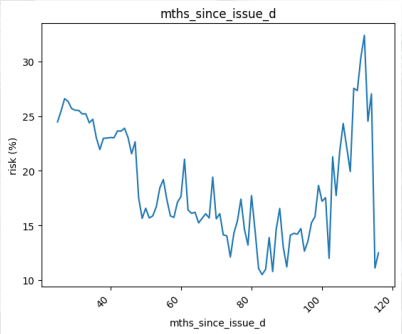


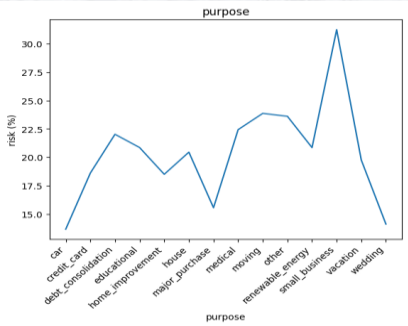
Exploratory Data Analysis

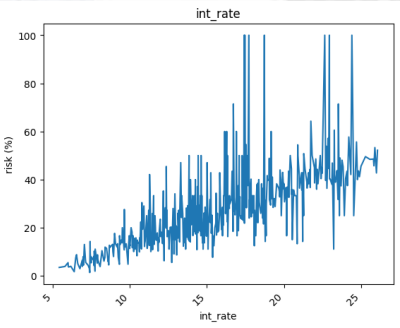












Exploratory Data Analysis

- **Lower credit grades (e.g., G) are associated with higher default risks.** Borrowers in these grades typically have weaker financial standing, as reflected in their credit assessments.
- **Longer loan terms, involving more installments, are correlated with a higher likelihood of bad credit.** Borrowers may struggle with long-term repayment commitments, especially if financial circumstances change.
- **A higher number of delinquent accounts directly increases the likelihood of default.** These accounts suggest past financial mismanagement or challenges in meeting financial obligations.
- **An increased number of credit checks within the last 6 months is a strong predictor of higher default risk.** This behavior often indicates financial distress or a borrower seeking multiple credit lines, potentially signaling overextension.

Exploratory Data Analysis

- **Borrowers who own or rent a home show a higher risk for bad loans. This is an interesting finding,** especially as borrowers without homes tend to exhibit a lower risk.

- **Loans with terms exceeding 100 months show a trend of increasing risk over time.** Long-term loans often face more uncertainty, as borrowers' financial circumstances can change significantly over extended periods.

- **Loans used for small businesses are the most at risk.** Small businesses typically face greater financial volatility. They are more susceptible to market shifts, economic downturns, and cash flow problems, making loans for small businesses inherently riskier.

- **Borrowers with a longer borrowing history often demonstrate reliable repayment behavior and build a positive credit track record.** However, this isn't always a perfect measure, as other factors like changes in income or economic conditions can still affect risk.

Exploratory Data Analysis

- **Debt in the whole (W) market tends to carry slightly higher risk than debt in the fractional (F) market.** This may be due to different investor profiles or underwriting standards across these markets.
- **Borrowers with more than 2 bills per year (excluding medical bills) face a higher risk of default.** This could indicate difficulty in managing day-to-day expenses, leading to financial strain.
- **There is a general trend that higher interest rates correlate with increased default risk.** Borrowers paying higher rates may already be considered higher risk, thus having a greater likelihood of defaulting.

Exploratory Data Analysis

Loan Amount, Grade, Loan Status

grade	A	B	C	D	E	F	G
loan_stats							
0	12148.124609	12343.340680	13232.306342	13946.722263	17123.724490	18482.620968	20327.254283
1	12161.520538	12563.370202	13686.239919	14666.521776	17679.200172	19072.932331	20498.538961

Annual Income, Grade, Loan Status

grade	A	B	C	D	E	F	G
loan_stats							
0	79161.655939	71886.931332	71160.443073	71333.372768	77673.799868	80197.316820	93441.470667
1	67561.592054	63675.745411	63143.778140	64179.295047	69351.350902	71360.196087	82080.278820

Revolving Balance, Grade, Loan Status

grade	A	B	C	D	E	F	G
loan_stats							
0	15011.325403	15081.987178	14969.181505	15243.415555	16872.074542	16923.059908	21575.368801
1	14807.946496	14612.173622	14781.833635	14841.688893	16273.146611	16755.600251	18005.908009

Home Ownership vs Status

home_ownership	ANY	MORTGAGE	NONE	OTHER	OWN	RENT
loan_stats						
0	1	96197	40	141	15917	78795
1	0	22640	8	38	4515	23767

Exploratory Data Analysis

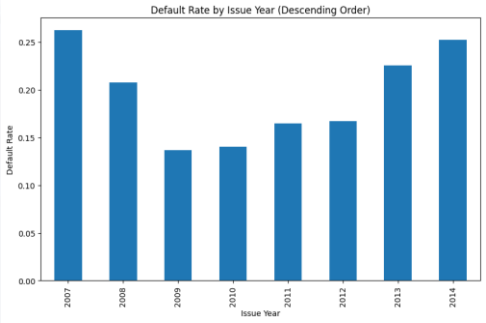
- **Borrowers with mortgages or rental statuses are more likely to take out loans.** Mortgage holders may be using debt to manage high financial obligations, while renters might face less stable financial situations.

- **Individuals with lower credit grades often carry higher revolving balances, indicating a greater reliance on loans and an increased likelihood of carrying unpaid debt month to month.** This behavior suggests higher financial risk, as larger revolving balances can lead to greater interest accrual and potential difficulty in managing debt.

- **Borrowers in lower credit grades tend to take on larger debts, further increasing their risk of default.** This trend suggests that lower creditworthiness does not always curb loan sizes, potentially exacerbating risk.

- **Borrowers with bad loans typically have lower average annual incomes, reflecting a strong link between income and credit risk.** Lower-income individuals may struggle more to meet debt obligations, leading to higher default rates.

Exploratory Data Analysis



issue_year	0	1	default_rate
2007	445	158	0.262023
2008	1897	496	0.207271
2009	4558	723	0.136906
2010	10771	1759	0.140383
2011	16509	3259	0.164862
2012	41589	8334	0.166937
2013	57693	16798	0.225504
2014	57629	19441	0.252251

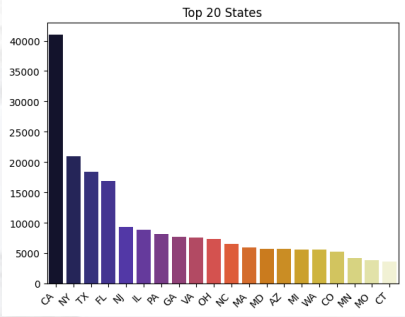
The high default rate of 26% in 2007 can likely be attributed to the onset of the global financial crisis, which started in that year. The housing market collapse, widespread foreclosures, and tightening credit conditions caused significant financial strain, leading to a spike in defaults. **In 2008, the default rate decreased to 20.7%, which, while still high, reflects the ongoing effects of the crisis as governments and financial institutions implemented emergency measures**, including bailouts and loan modifications, to stabilize the economy.

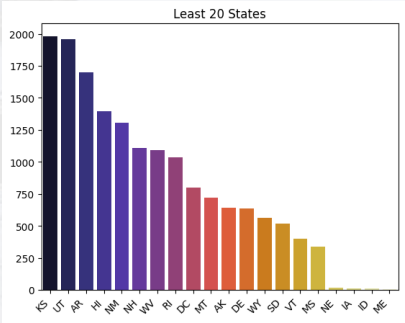
Exploratory Data Analysis

From 2009 to 2012, the default rate continued to decrease to between 13-16%. This decline likely reflects the gradual economic recovery following the recession, as markets stabilized and job growth slowly improved. Additionally, lending practices became stricter, reducing the number of high-risk loans and improving overall loan quality. The Dodd-Frank Act and other post-crisis financial regulations may have played a key role in reducing risky lending, leading to a much healthier credit environment.

However, the rate rose again to 22-25% during 2013-2014. This increase can be attributed to heightened economic uncertainty, driven by the U.S. debt ceiling crisis, global market volatility, and the Federal Reserve's tapering of its quantitative easing program. The political standoff over the debt ceiling created fears of a government default, while concerns about slower growth in emerging markets and the Eurozone debt crisis intensified market instability. Additionally, the Fed's decision to reduce its bond-buying program led to rising interest rates, further straining borrowers. These combined factors made it difficult for businesses and borrowers to manage financial pressures, resulting in higher default rates during this period.

Exploratory Data Analysis





- **California (CA) takes on the most debt, which can be attributed to its large population and high cost of living.** New York (NY), Texas (TX), and Florida (FL) also have high debt, though at lower levels, reflecting regional economic diversity and varying cost pressures. For instance, NY's financial hub and TX's rapid population growth both contribute to higher levels of borrowing.

- **States like Nebraska (NE), Iowa (IA), Idaho (ID), and Maine (ME) report the lowest levels of debt, with fewer than 500 accounts in debt.** These states' smaller populations, lower living costs, and more conservative attitudes toward debt contribute to less reliance on credit.

Exploratory Data Analysis

```
Top 5 highest bad loan ratio states:
addr_state
NE    0.571429
MS    0.312500
TN    0.290901
IN    0.269081
NV    0.249094
Name: bad_loan_ratio, dtype: float64

Top 5 lowest bad loan ratio states:
addr_state
NH    0.166517
WY    0.147687
DC    0.121250
ID    0.111111
ME         NaN
Name: bad_loan_ratio, dtype: float64
```

- **The higher default rates in states like Nebraska (NE), Mississippi (MS), Tennessee (TN), Indiana (IN), and Nevada (NV) likely reflect economic challenges** such as lower incomes, higher unemployment, or market instability. Nebraska's particularly high default ratio of 0.57% may be linked to localized issues like industry downturns or agricultural instability.

- **States like New Hampshire (NH), Wyoming (WY), and Washington, D.C. (DC) have low bad loan rates,** indicating stronger financial stability or conservative borrowing habits. Maine (ME) stands out with no recorded bad loans, possibly due to a low-risk borrower base or effective lending practices, influenced by smaller populations and fewer risky loans.

Actionable Steps

- **Borrowers with lower credit grades face higher default risk due to larger debts.** To manage this, stricter approval, higher interest rates, borrowing limits, and early warning systems should be implemented.

- **Longer loan terms are linked to higher default risk, as borrowers may face financial instability over time. Consider limiting long-term loans or introducing tighter conditions** (e.g., higher down payments, interest rates) for high-risk borrowers to reduce the likelihood of default.

- **Borrowers with multiple delinquent accounts are at a high risk of default. Implement stricter credit checks** for these individuals, and consider offering financial counseling or restructuring options for borrowers struggling with multiple delinquencies.

- **Bad loans tend to involve larger amounts and higher interest rates. Implement stricter lending policies for borrowers requesting large loans**, particularly in lower credit grades, and assess their ability to repay before approval.

Actionable Steps

- **Develop tailored financial products or offer budget management tools for borrowers who own or rent homes**, as they present a higher risk for bad loans due to larger financial obligations.

- **Offer more comprehensive risk assessments for loans used in small businesses**, possibly tightening lending terms for this segment due to their susceptibility to financial volatility.

- **Borrowers with lower incomes are more likely to default. Income should be a key factor in credit scoring models, and lenders should focus on offering loans that are within a borrower's financial capacity**, potentially through income verification or offering smaller loan sizes to lower-income applicants.

- **Lenders in states with high borrowing levels should adjust loan limits and interest rates to align with regional economic conditions, while states with lower debt may require more conservative lending practices**, focusing on smaller loans. Additionally, in states with high default rates, stricter underwriting criteria and enhanced risk assessments are essential to mitigate risks caused by localized economic stress.

Additional Data Cleaning + Preprocessing

Removing Columns

```
#Removing unnecessary data
dropped = ['emp_length', 'emp_title', 'mths_since_issue_d', 'sub_grade', 'issue_d', 'issue_year']
df.drop(columns = dropped, axis=1, inplace=True)
```

Additional irrelevant columns are removed. For example, emp_title and emp_length are excluded due to too many unique values and limited relevance to credit risk. Sub-grade is dropped in favor of grade, which simplifies the model by reducing category complexity. Also, columns like mths_since_issue_d and issue_year are removed because they contain data unavailable at the time of loan approval.

Addressing Null Values

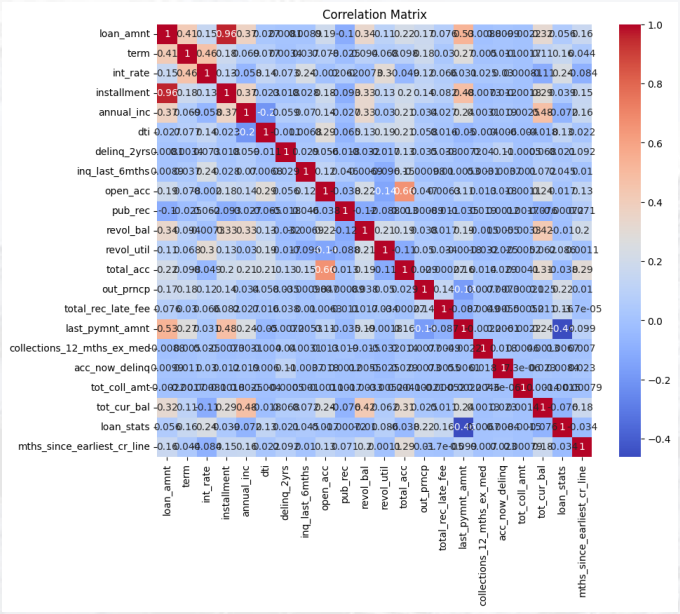
delinq_2yrs	29
inq_last_6mths	29
open_acc	29
pub_rec	29
revol_util	234
total_acc	29
initial_list_status	0
last_pymnt_amnt	0
collections_12_mths_ex_med	145
acc_now_delinq	29
tot_coll_amt	66689
tot_cur_bal	66689
loan_stats	0
mths_since_earliest_cr_line	29

```
df.dropna(inplace=True)
```

```
df['open_acc'] = df['open_acc'].astype(int)
df['pub_rec'] = df['pub_rec'].astype(int)
df['total_acc'] = df['total_acc'].astype(int)
df['acc_now_delinq'] = df['acc_now_delinq'].astype(int)
```

Rows with excessive missing values are dropped to ensure the integrity of the results. Additionally, some columns are converted to the appropriate data type, which is integer.

Feature Correlations




```python
# Find columns with correlation above 0.7
upper_tri = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(bool))
to_drop = [column for column in upper_tri.columns if any(upper_tri[column] > 0.7)]
print(to_drop)

# Drop the highly correlated columns
df.drop(to_drop, axis=1, inplace=True)
```

```
['installment']
```

Features with a correlation above 0.7 are removed to prevent overfitting. For example, the loan amount shows a high correlation with installment amounts, making one of these variables redundant in the model.

Removing Outliers

```
outlier = ['annual_inc', 'last_pymnt_amnt','tot_coll_amt','tot_cur_bal']
```

```
print(f'Count of rows before removing outlier: {len(df)}')
for i in outlier:
  df = subset_by_iqr(df, i)
print(f'Count of rows after removing outlier: {len(df)}')
```

```
Count of rows before removing outlier: 175256
Count of rows after removing outlier: 141002
```

Outlier data is removed to prevent it from skewing the model's performance.

Labeling Categorical Data

```
# One-hot encoding for categorical features
cat_features = df.select_dtypes(include=['object']).columns
df = pd.get_dummies(df, columns=cat_features, drop_first=True)
```

Categorical data is one-hot encoded to ensure the model can process it effectively, as machine learning models cannot handle string-type data.


id/x partners
Rakamin
Academy
Modeling

Data Split

```
#Defining x and y
x = df.drop(columns=['loan_stats'], axis = 1)
y = df['loan_stats']

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.2, random_state = 42)
```

The data is split into independent variables (x) and the target variable (y, representing loan status). An 80/20 train-test split is applied to the dataset.

Data Balancing

```
from imblearn.over_sampling import SMOTE

smote = SMOTE(random_state=42)
x_train, y_train = smote.fit_resample(x_train, y_train)
```

Due to class imbalance, the training data is oversampled using SMOTE to ensure balanced representation. Only the training data is oversampled to preserve the original distribution of the test data and avoid overfitting.

Modeling

```
from sklearn.ensemble import RandomForestClassifier
rf = RandomForestClassifier(class_weight={0: 1, 1: 10})

#Training the model
rf.fit(x_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(class_weight={0: 1, 1: 10})
```

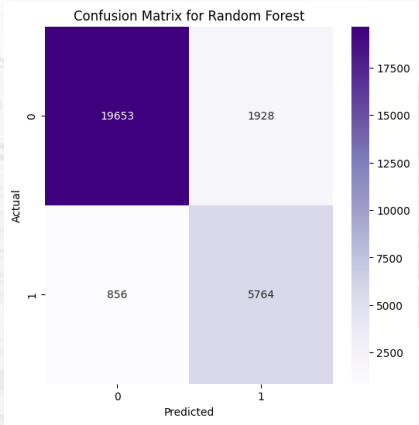
A Random Forest model is used with adjusted class weights to assign a higher cost to misclassifying bad loans, as these pose a greater financial risk.

Evaluation

	precision	recall	f1-score	support
0	0.96	0.91	0.93	21581
1	0.75	0.87	0.81	6620
accuracy			0.90	28201
macro avg	0.85	0.89	0.87	28201
weighted avg	0.91	0.90	0.90	28201

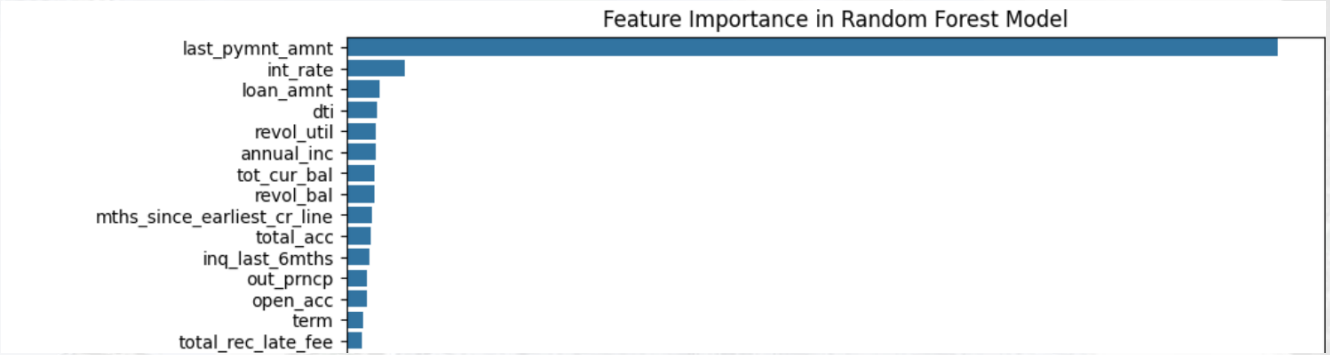
```
#Check model performance using auc score
roc_auc_score(y_test, y_pred)*100
```

```
89.06785102874564
```



The model performs well in detecting bad loans, even with imbalanced data. The model's AUC score reaches 89%, indicating strong predictive performance and reliability in identifying high-risk loans.

Feature Importance



The chart above illustrates the feature importance, highlighting how each factor contributes to the Random Forest model's ability to predict credit risk. The most significant feature is last_pymnt_amnt, indicating that the last payment amount plays a key role in determining credit risk. Int_rate (interest rate) is also a critical feature. It's important to note that the chart excludes categorical features, which have already been one-hot encoded.

Thank You!

Connect with Me

LinkedIn GitHub