IBM **SkillsBuild**

Loan Defaulter Prediction System

Team ArtiSci'09

Start Slide

IBM **SkillsBuild**

Meet With Amazing Team

Lokesh Patra

D Vamsi Krishna

Soumya Khuntia

IBM **SkillsBuild**

Meet With Amazing Team

Subham Pratik Das

Indigibilli Harshit

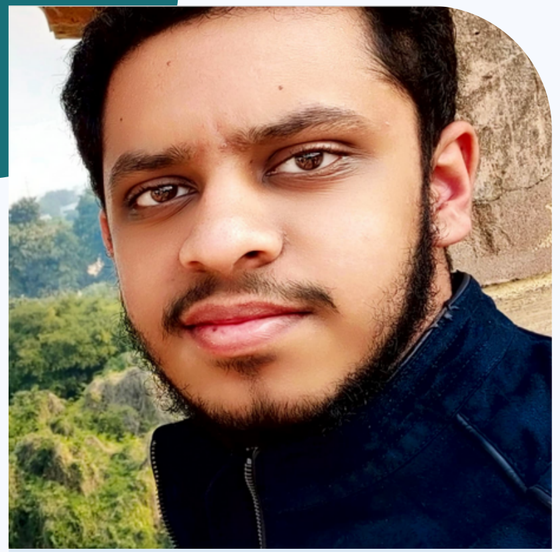
Priyanka Mohapatra

IBM **SkillsBuild**

Meet With Amazing Team

Rudra Narayan Behera

Aakash Kumar Mahato



Ishan Kumar

IBM **SkillsBuild**

Bank Loan

A "bank loan" is a financial arrangement where a borrower borrows a specific amount of money from a bank and agrees to repay it over a set period of time, usually with interest.

Loan Defaulter Prediction System

The "Loan Defaulter Prediction System" is an innovative financial tool that utilizes advanced data analysis and machine learning algorithms to assess the creditworthiness of borrowers. By analyzing historical data and various factors, it aims to accurately predict the likelihood of a borrower defaulting on a loan, assisting lenders in making informed decisions.

- ✓ **Asset Checking**
- ✓ **Market Services**
- ✓ **Corporate Trust**
- ✓ **Credits Market**

IBM **SkillsBuild**

Summary

The "**Loan Defaulter Prediction System**" project is a robust data-driven solution that utilizes the **Random Forest** machine learning algorithm to accurately predict loan defaulters. It processes and analyzes a vast dataset of **historical loan information**, extracting **pertinent features** to train an ensemble of **decision trees**.

Through this ensemble approach, the system achieves **higher accuracy** and minimizes **overfitting**. By providing lenders with **actionable insights**, it enables them to **proactively identify** and mitigate **potential default risks**, optimizing the lending process, and effectively **minimizing financial losses**.

IBM **SkillsBuild**

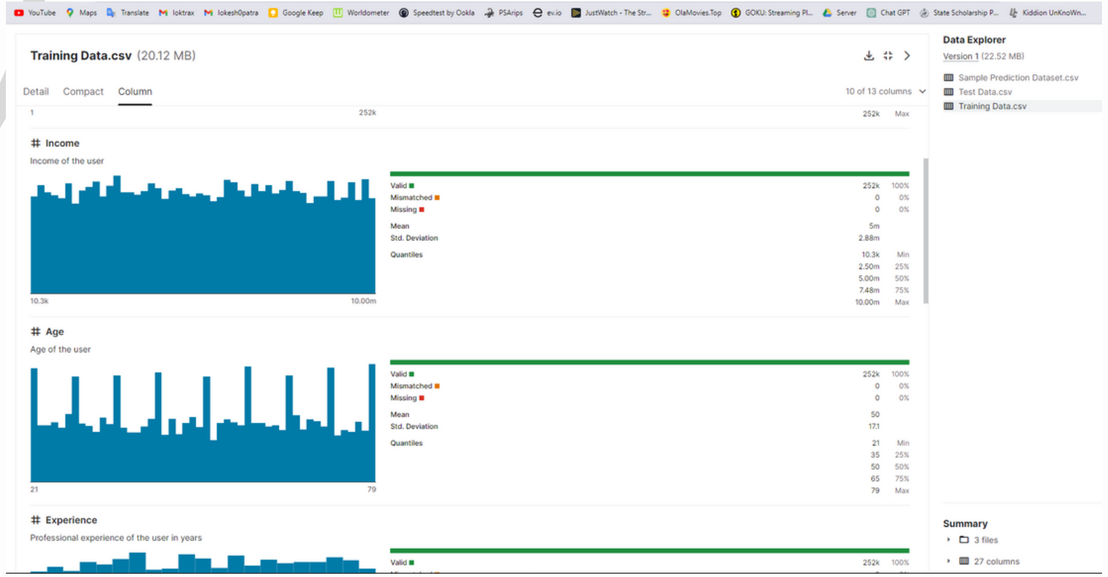
What is its purpose ?

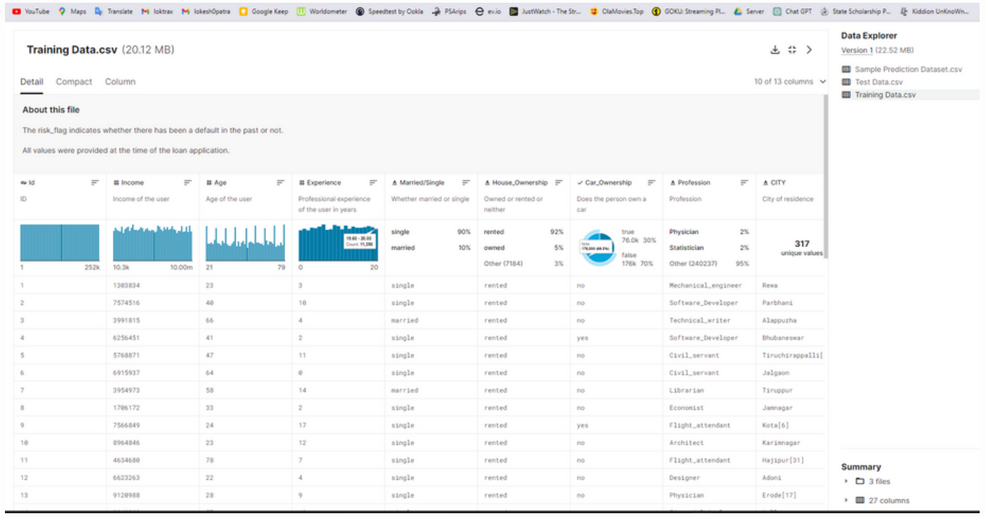
The system identifies borrowers who are at **higher risk of defaulting** on their loan payments. This predictive insight empowers lenders to take **proactive measures**, such as **adjusting loan terms** or declining **high-risk applications**, to **minimize financial losses** and improve the **overall efficiency** and profitability of their lending operations. Ultimately, the project's primary goal is to **enhance risk management**, streamline lending processes, and ensure **responsible** and **sustainable lending practices**.

IBM **SkillsBuild**

Data Source

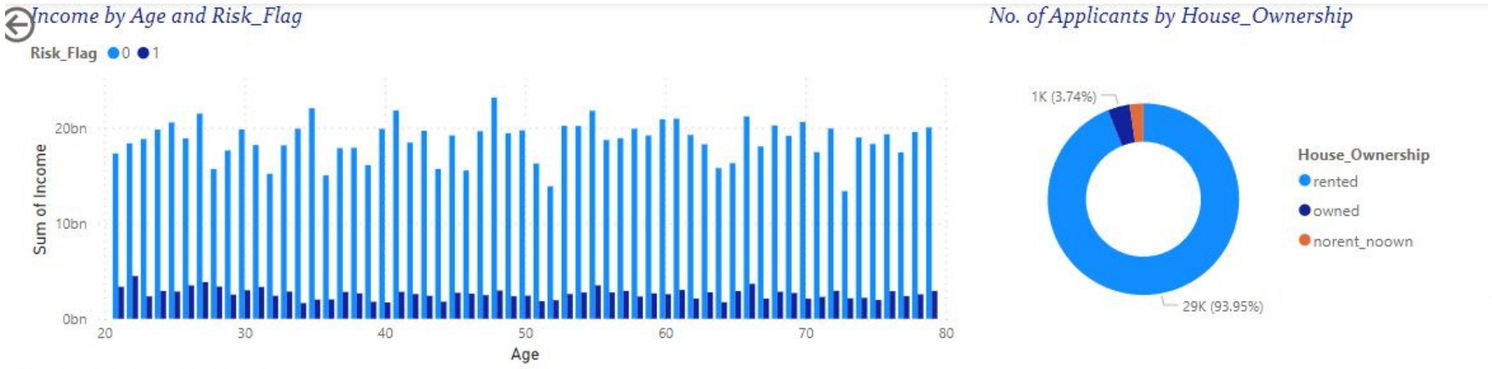
- https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior?resource=download





IBM **SkillsBuild**

Visualisation

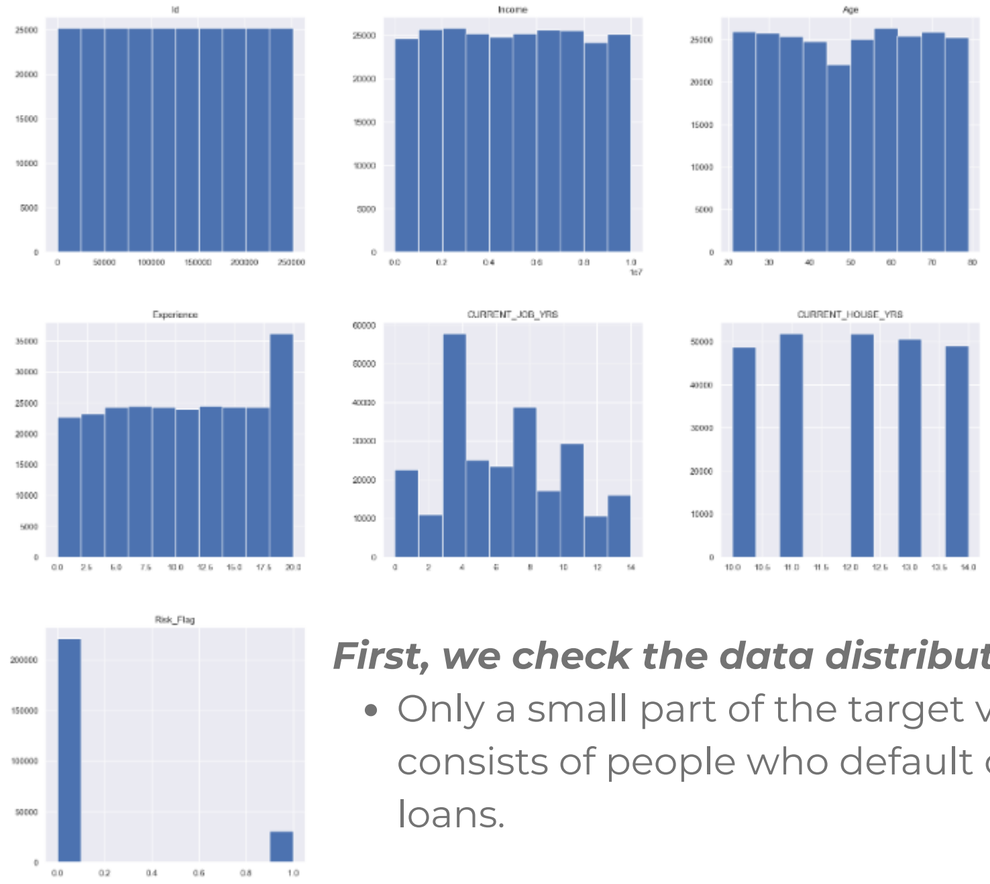


- **Clustered BarChart:** People with lower income are marked at risk of defaulting, irrespective of their age.
- **Doughnut Chart:** Highest no. Of applicants have rented house.

- **TreeMap:** Uttar Pradesh has the highest no. of possible defaulters followed by West Bengal and Andhra Pradesh. UP has 3343 no. of defaulters according to the dataset.

IBM **SkillsBuild**

Visualisation

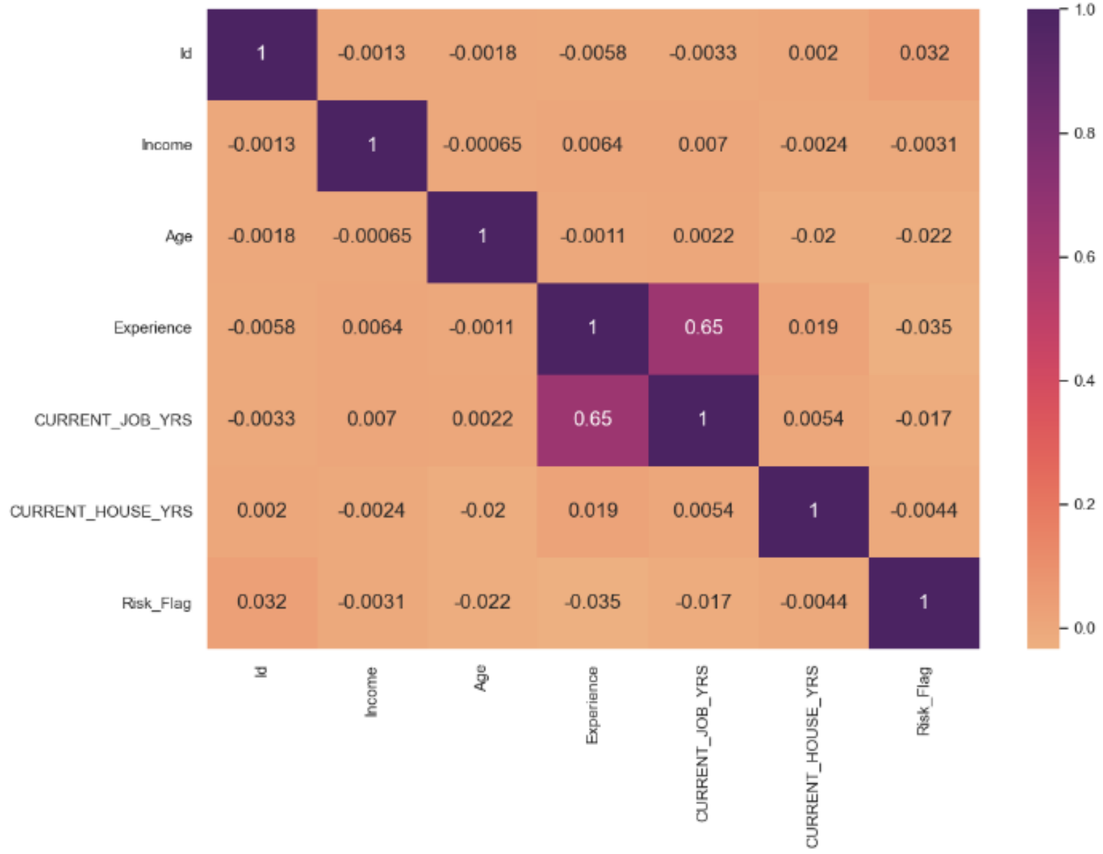


First, we check the data distribution.

- Only a small part of the target variable consists of people who default on loans.

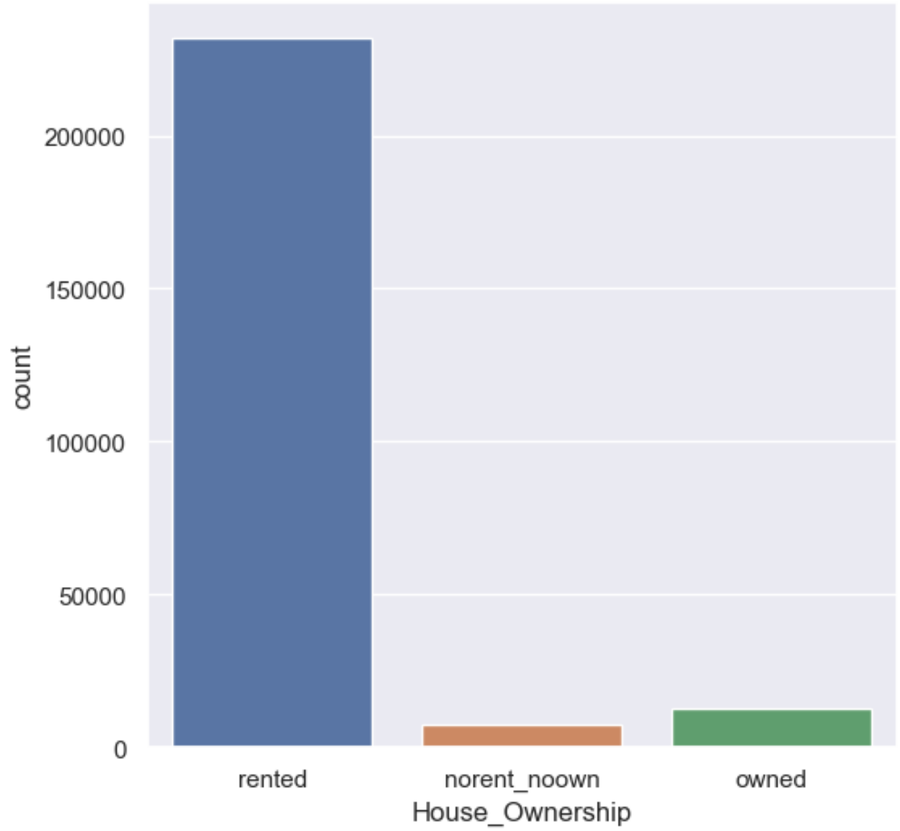
Correlation Matrix

- We notice there is a fairly strong correlation between experience and current job years, which is self justified.

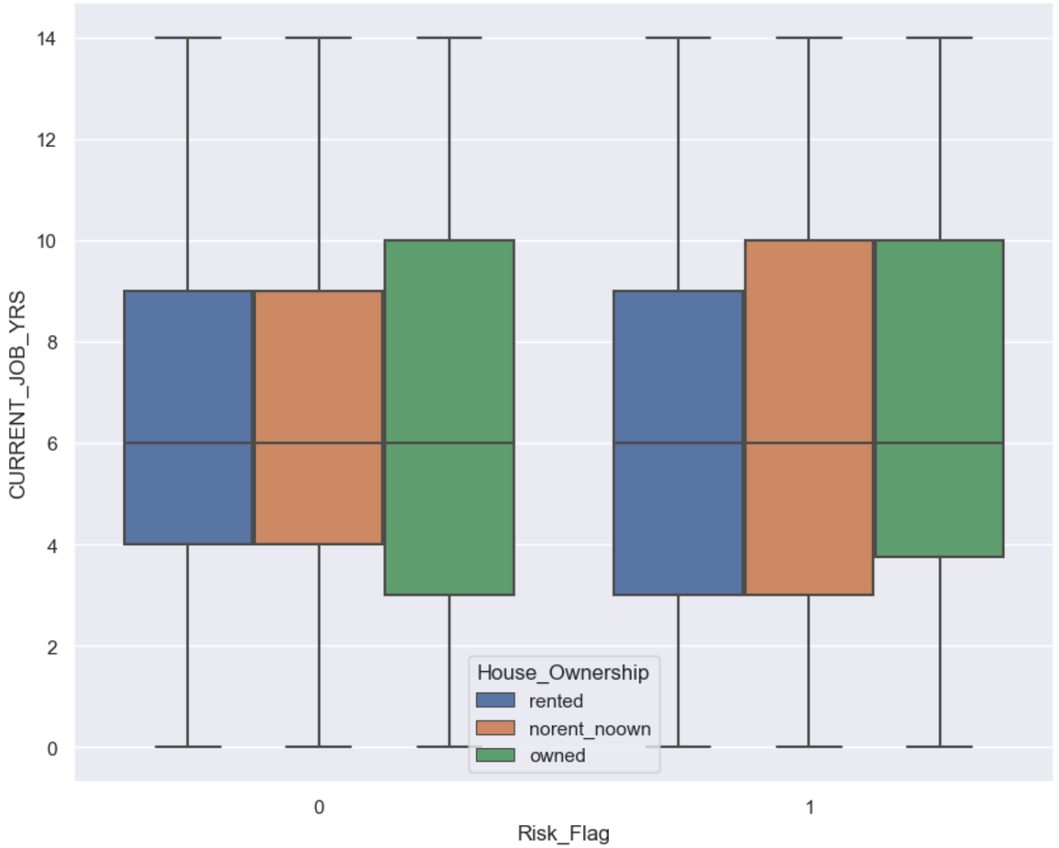


IBM **SkillsBuild**

Visualisation



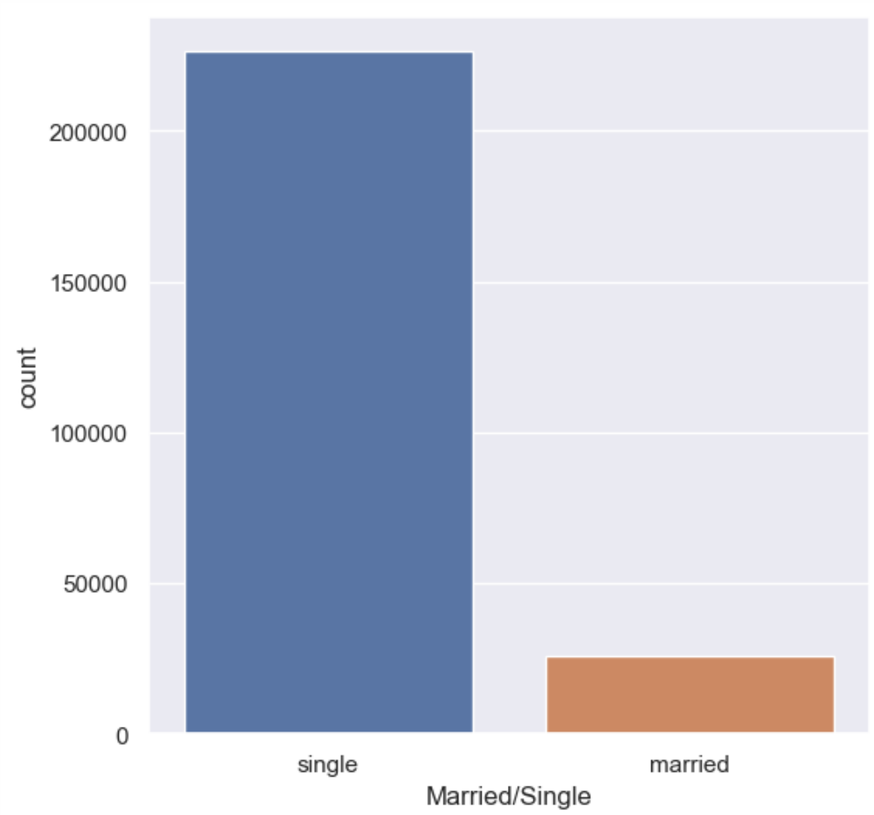
- Majority of the loan applied people have a rented accommodation.



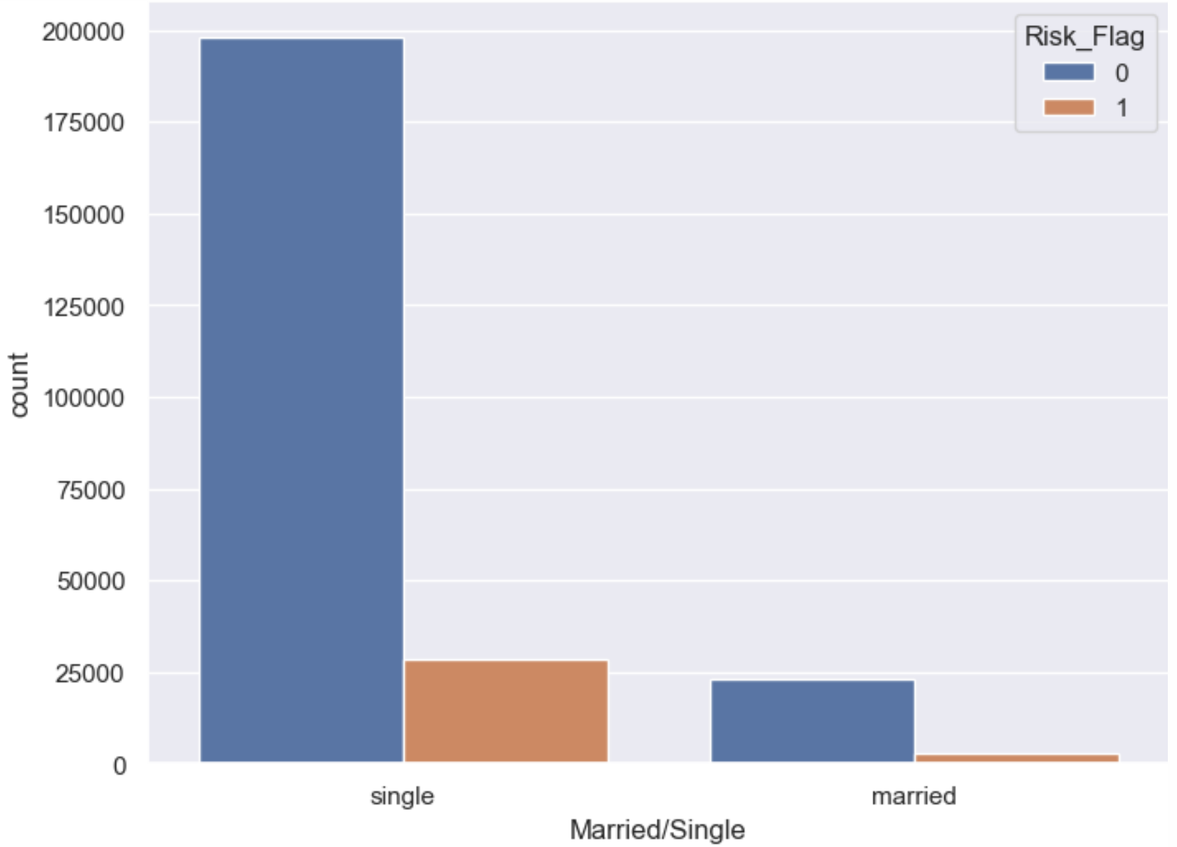
- We see that higher no. of possible defaulters neither have a rented or own house whereas most of the people who are unlikely to default their loan have their own house.

IBM **SkillsBuild**

Visualisation



- The majority of the applicants are single.



- And higher no. of possible defaulters are single.

IBM **SkillsBuild**

Prediction

In [7]:
```
from sklearn.preprocessing import LabelEncoder

label_encoder = LabelEncoder()
for col in ['Married/Single', 'Car_Ownership']:
    data[col] = label_encoder.fit_transform(data[col])
```

In [8]:
```
onehot_encoder = OneHotEncoder(sparse = False)
data['House_Ownership'] = onehot_encoder.fit_transform(data['House_Ownership'].values.reshape(-1, 1) )
```

In [9]:
```
high_card_features = ['Profession', 'CITY', 'STATE']
count_encoder = ce.CountEncoder()
# Transform the features, rename the columns with the _count suffix, and join to dataframe
count_encoded = count_encoder.fit_transform( data[high_card_features] )
data = data.join(count_encoded.add_suffix("_count"))
data= data.drop(labels=['Profession', 'CITY', 'STATE'], axis=1)
```

In [10]:
```
x = data.drop("Risk_Flag", axis=1)
y = data["Risk_Flag"]
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, stratify = y, random_state = 7)
```

In [13]:
```
from sklearn.ensemble import RandomForestClassifier
from imblearn.over_sampling import SMOTE
from imblearn.pipeline import Pipeline
rf_clf = RandomForestClassifier(criterion='gini', bootstrap=True, random_state=100)
smote_sampler = SMOTE(random_state=9)
pipeline = Pipeline(steps = [['smote', smote_sampler],
                             ['classifier', rf_clf]])
pipeline.fit(x_train, y_train)
y_pred = pipeline.predict(x_test)
```

```
y_pred[0:5]
```

```
array([1, 0, 1, 0, 0], dtype=int64)
```

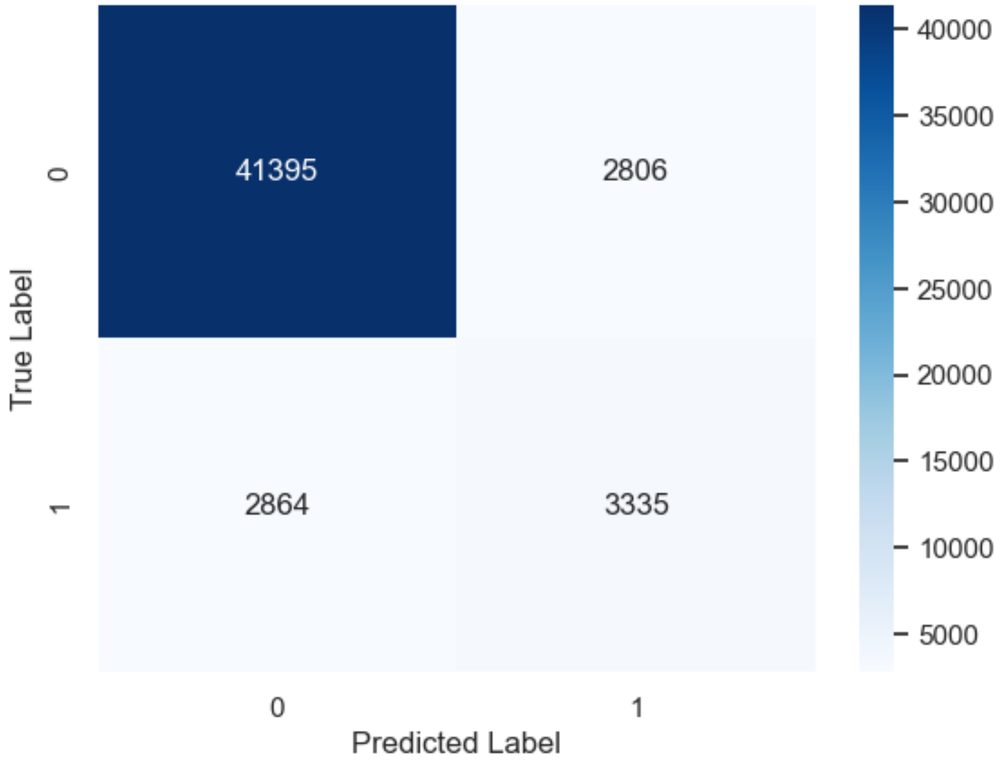
- Sample prediction has been done based on the first five observations of testing data.

IBM **SkillsBuild**

Evaluation

```
-----------------------TEST SCORES--------------------
Recall: 53.799
Precision: 54.3071
F1-Score: 54.0519
Accuracy score: 88.75
```

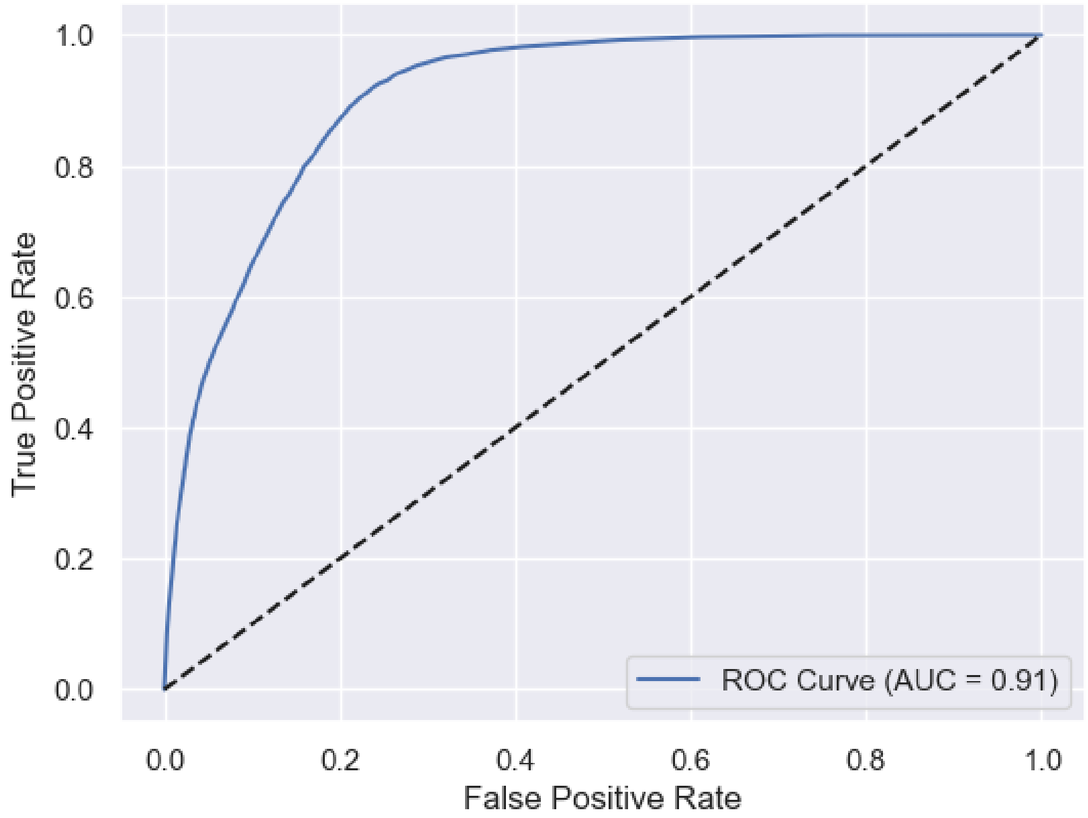
- The **precision of 54.31%** indicates that when the model predicts a sample as a **"yes"** (*default, positive class*), it is correct approximately 54.31% of the time. ***Higher precision*** means the model has a **low rate of false positives**.
- The **recall of 53.80%** means that the model correctly identifies approximately 53.80% of the **actual positive samples** (*defaults*).
- The **F1 score of 54.05%** is the ***harmonic mean*** of **precision and recall** and provides **a balanced measure** between them.
- The **accuracy of 88.75%** indicates the **overall correctness** of the model's predictions.



- Based on **the evaluation metrics** , we can plot the ***confusion matrix***
- The Confusion Matrix shows the **performance** of the **Random Forest model** for **binary classification**.
- It correctly predicted **41,395 positive instances** (*True Positives*) and **3,335 negative instances** (*True Negatives*).
- However, it made **2,806 incorrect positive predictions** (*False Positives*) and **2,864 incorrect negative predictions** (*False Negatives*).

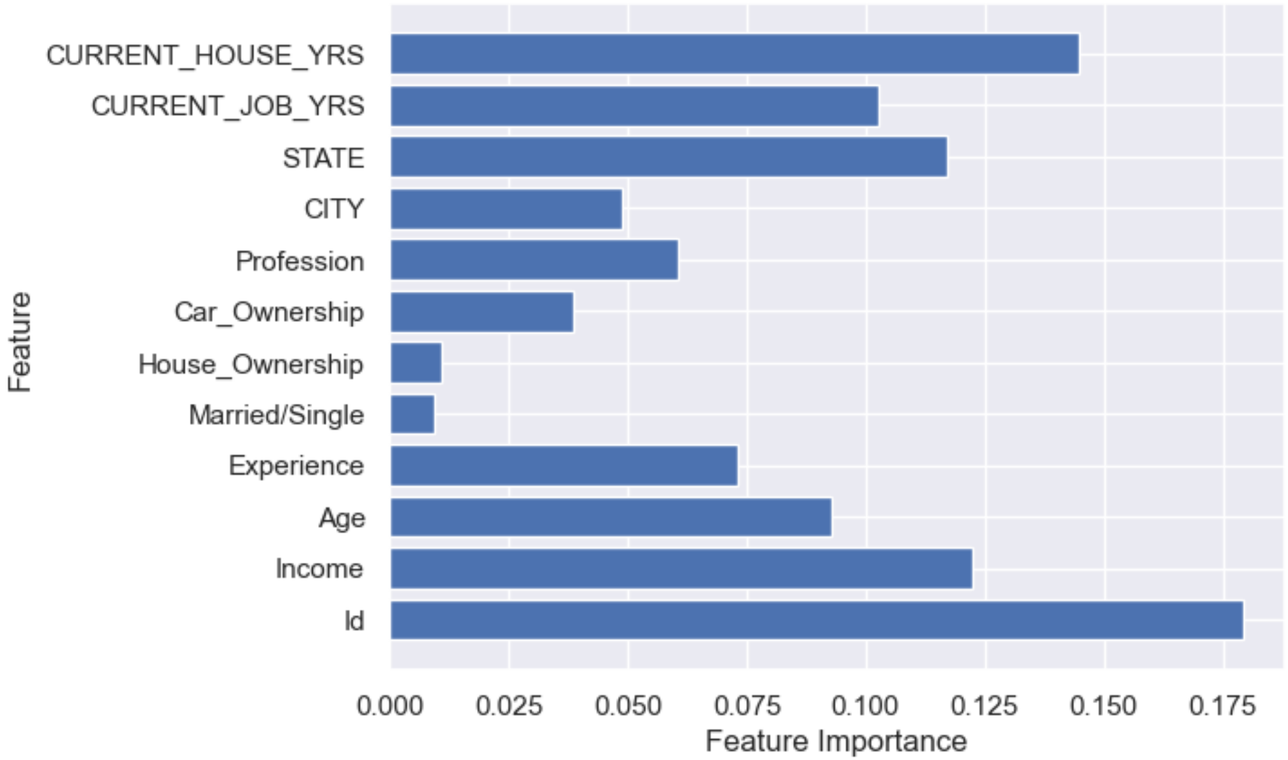
IBM **SkillsBuild**

Evaluation



ROC Curve

- High AUC indicates good model discrimination on *imbalanced data.*



- Relatively higher scores of current house years, state and income signifies the high influence of these features in the prediction.

IBM **SkillsBuild**

Significance

The "Loan Defaulter Prediction System" is essential due to several compelling reasons:

1. **Risk Mitigation**: Lending institutions face significant risks when disbursing loans, as there is always a chance of borrowers defaulting on their payments. By implementing a robust prediction system, lenders can proactively identify borrowers who are more likely to default, allowing them to take preventive measures or adjust the loan terms to reduce potential losses.

2. **Improved Decision Making**: Traditional lending decisions rely heavily on manual assessments and historical data analysis, which may not be efficient in identifying complex patterns. The prediction system leverages sophisticated machine learning algorithms, enabling lenders to make data-driven decisions based on a comprehensive analysis of numerous variables and borrower behavior.

3. **Cost Reduction**: Loan defaults result in significant financial losses for lenders, including the costs associated with debt collection, legal proceedings, and recovery efforts. By accurately predicting potential defaulters, lenders can avoid extending loans to high-risk applicants, thus minimizing these costs and improving their overall financial performance.

4. **Enhanced Customer Experience**: A reliable prediction system ensures that loans are extended to borrowers who have a higher likelihood of repaying on time. This reduces the chances of burdening borrowers with loans they cannot manage, leading to improved customer satisfaction and increased trust in the lending institution.

IBM **SkillsBuild**

Significance

5. **Regulatory Compliance**: Many financial institutions must comply with regulatory requirements that focus on responsible lending practices. Implementing a Loan Defaulter Prediction System helps lenders adhere to these regulations by ensuring fair and unbiased decision-making processes based on objective data analysis.

6. **Scalability and Efficiency**: As the lending industry grows and the number of loan applicants increases, manual evaluation becomes increasingly impractical. The prediction system's automated approach allows lenders to scale their operations efficiently, serving a larger customer base without compromising on the quality of loan assessments.

7. **Predictive Insights**: The system not only identifies potential defaulters but also provides valuable insights into borrower behavior and risk factors. Lenders can use this information to refine their loan products, optimize interest rates, and develop customized financial solutions tailored to different customer segments.

IBM **SkillsBuild**

Industrial Use Case

The practical industrial use of the "Loan Defaulter Prediction System" project would be in the financial and banking sectors. Lending institutions, such as banks, credit unions, and online lenders, can implement this system to enhance their loan approval process and risk management strategies. Some specific industrial use cases include:

1. **Loan Application Screening:** When borrowers apply for loans, the prediction system can assess their creditworthiness based on historical loan data and relevant features. Lenders can use this analysis to approve or reject loan applications, ensuring loans are extended to customers who have a higher likelihood of repaying on time.
2. **Risk Assessment:** The system can evaluate the risk associated with each loan application and assign a risk score to applicants. Lending institutions can use these scores to prioritize their loan portfolio and allocate resources effectively to manage potential default risks.

3. **Portfolio Management:** With insights into potential default risks, lenders can make data-driven decisions regarding their existing loan portfolio. They can identify high-risk accounts and implement proactive measures, such as loan restructuring or collections efforts, to mitigate losses.
4. **Interest Rate Optimization:** By understanding the credit risk of borrowers, lenders can adjust interest rates according to the level of risk associated with each loan. Lower-risk applicants may qualify for lower interest rates, while higher-risk applicants may receive loans with higher interest rates.

IBM **SkillsBuild**

Industrial Use Case

5. **Compliance and Responsible Lending:** The system can help lending institutions comply with regulatory requirements and responsible lending practices. It ensures that loan decisions are based on objective data analysis, minimizing bias and promoting fair lending practices.
6. **Customer Segmentation:** The prediction system can segment customers based on their credit risk profiles. Lenders can use this information to offer targeted loan products and personalized financial solutions to different customer segments, increasing customer satisfaction and retention.
7. **Fraud Detection:** In addition to predicting loan defaulters, the system can also help detect fraudulent loan applications by analyzing suspicious patterns and anomalies in the data.

Conclusion

In conclusion, the "Loan Defaulter Prediction System" is a critical tool for modern lending institutions. It leverages advanced data analysis and machine learning to enhance decision-making, reduce financial risks, improve customer satisfaction, and ensure compliance with regulatory standards. By embracing such a system, lenders can achieve greater efficiency, profitability, and long-term sustainability in their lending operations.

By implementing this system, financial institutions can optimize their loan approval process, reduce financial losses, improve customer experience, and maintain a healthy and profitable loan portfolio. This advanced data-driven approach revolutionizes the way lenders manage risk and make lending decisions, leading to more efficient and sustainable lending practices in the financial industry.

IBM **SkillsBuild**

THANK YOU

Team
ArtiSci'09

THE ANALYTICAL ARCHITECTS