Loan Analysis & Prediction

Salem Arthur, Karl Hickel, Candice Biying Han, Kathy Yu-Hsin Lee

Agenda

- Introduction
- Project Process
- Data Description + Data Cleaning
- Data Visualization & Business Insight
- Machine Learning (KNN, Decision Tree, Random Forest)
- Conclusion

Introduction

- Founded in 2006 by Renaud Laplanche
- The world’s largest peer-to-peer lending platform.
- Provide loans of up to $40,000 for varying purposes
    - Ex. Credit card, small business, house, medical

Project Process

Project Objective

- Detailed insight into our customers
- If customer had difficulty paying back loan, what were the predicting factors of them defaulting.

Steps

Data Description

- Extensive information about customers personal info and loan status.
  - Ex. Purpose of loan, Combined household income, status of loan, occupation.
- Large dataset
  - 759338 rows × 72 columns

	id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	grade	sub_grade	home_ownership	annual_inc	loan_status	pymnt_plan
0	112435993	2300	2300	2300.0	36 months	12.62	C	C1	OWN	10000.0	Current	n
1	112290210	16000	16000	16000.0	60 months	12.62	C	C1	MORTGAGE	94000.0	Current	n
2	112436985	6025	6025	6025.0	36 months	15.05	C	C4	MORTGAGE	46350.0	Current	n
3	112439006	20400	20400	20400.0	36 months	9.44	B	B1	RENT	44000.0	Current	n
4	112438929	13000	13000	13000.0	36 months	11.99	B	B5	MORTGAGE	85000.0	Current	n
...	...	...	...	...	...	...	...	...	...	...	...	...
759333	65854936	6000	6000	6000.0	36 months	7.89	A	A5	OWN	38000.0	Current	n
759334	66055600	6000	6000	6000.0	36 months	9.17	B	B2	RENT	32640.0	Current	n
759335	66141895	14400	14400	14400.0	60 months	13.18	C	C3	RENT	47000.0	Late (16-30 days)	n
759336	65673209	34050	34050	34050.0	36 months	15.41	D	D1	MORTGAGE	87800.0	Current	n
759337	65744272	5000	5000	5000.0	36 months	11.22	B	B5	MORTGAGE	65000.0	Current	n

Data Cleaning/Feature building

- **Default Risk**

```
# Create the dictionary
default_dictionary ={'Current':'No Risk', 'Fully Paid':'No Risk', 'Charged Off':'Default',
                     'Late (31-120 days)':'High Risk', 'Late (16-30 days)':'Medium Risk',
                     'In Grace Period':'Medium Risk', 'Default':'Default'}

# Add a new column named 'default'
data['default_risk'] = data['loan_status'].map(default_dictionary)
```

- **Income Level**

```
#Use quantile to decide income level
data["annual_inc"].quantile([.25, .5, .75])
```

```
0.25    48000.0
0.50    67000.0
0.75    95000.0
```

```
data["income_level"].value_counts()
```

```
high        195890
mid_low     190822
low         188062
mid_high    184564
```

- **Refactoring**

```
In [25]: levels = {"No Risk":0, "Medium Risk":1,"High Risk":2,"Default":3}
         trainingData['default_risk'] = trainingData['default_risk'].map(levels)

In [26]: gradeLevels = {"A":0,"B":1,"C":2,"D":3,"E":4,"F":5,"G":6}
         trainingData["grade"] = trainingData["grade"].map(gradeLevels)
```

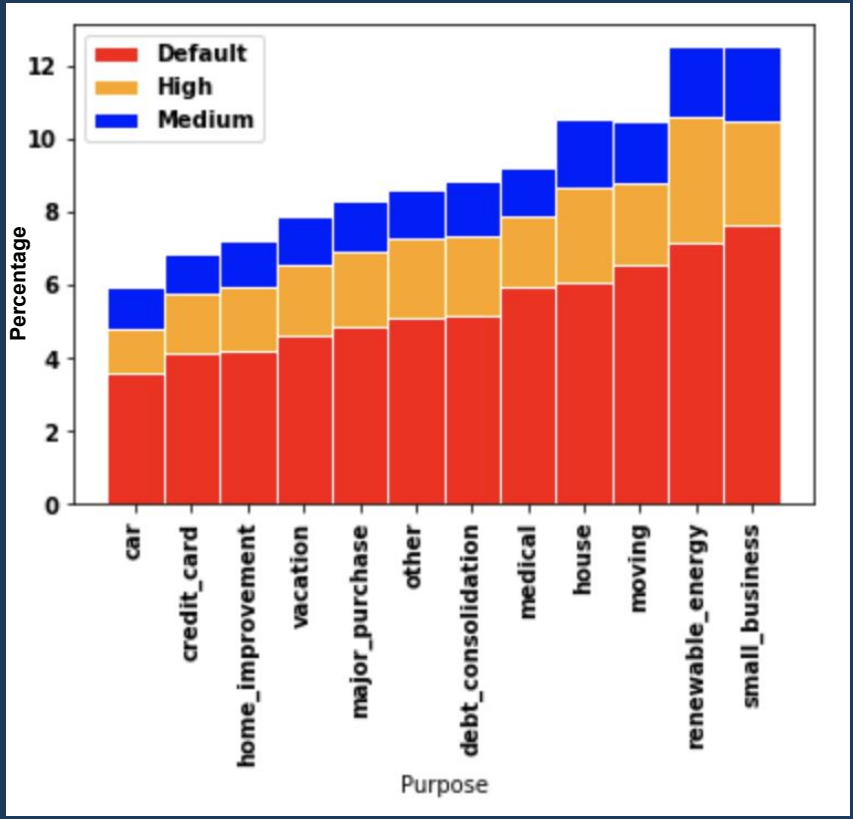
Return on Investment

```
# only consider the default and fully paid status

# defaulted loans
roi_default = 100*(data.loc[data['default_risk'] == 'Default', 'total_pymnt'].sum()
                   - data.loc[data['default_risk'] == 'Default', 'funded_amnt'].sum())
                    /data.loc[data['default_risk'] == 'Default', 'funded_amnt'].sum()
print('return on investment for defaulted loans: ',roi_default)

roi_paid = 100*(data.loc[data['loan_status'] == 'Fully Paid', 'total_pymnt'].sum()
                - data.loc[data['loan_status'] == 'Fully Paid', 'funded_amnt'].sum())
                /data.loc[data['loan_status'] == 'Fully Paid', 'funded_amnt'].sum()
print('return on investment for fully paid loans: ',roi_paid)
```

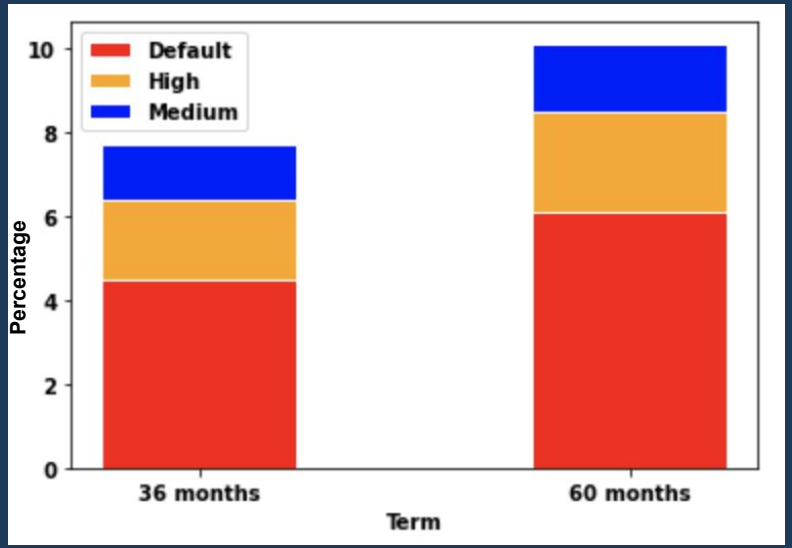
```
return on investment for defaulted loans:  -70.61041780037785
return on investment for fully paid loans:  8.489282767312927
```

- If a customer default, the company will lose 70.61% of the loan amount.
- If a customer fully paid, the company will earn 8.49% of the loan amount.

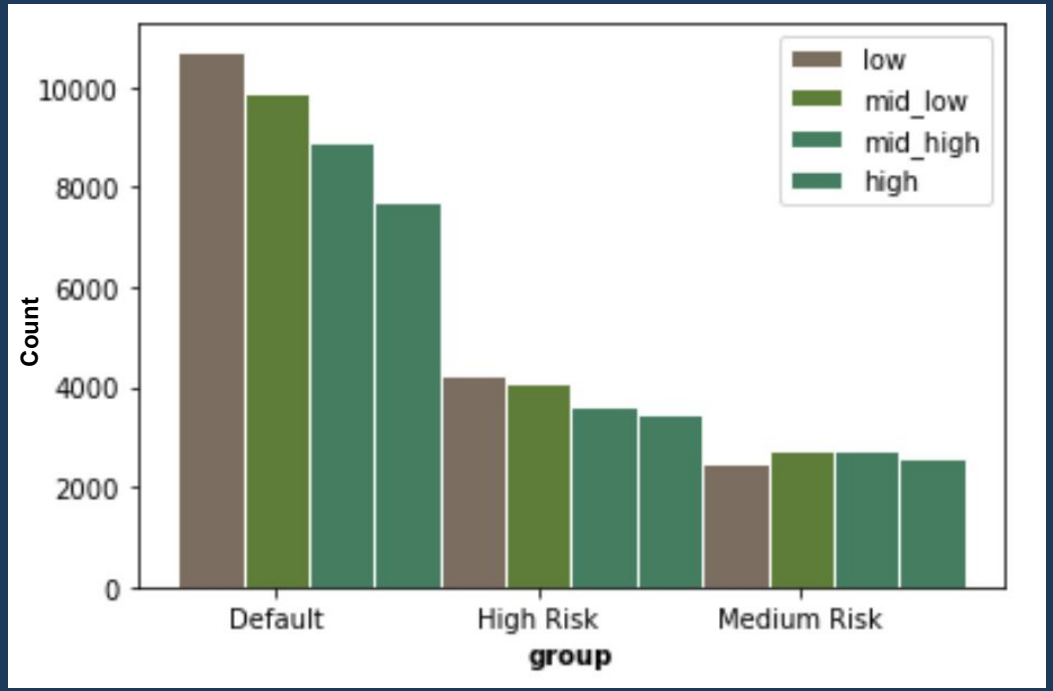
Default Risk



- Customer who get a loan for small business, renewable energy, and moving have higher default risk
- Customers with longer term have higher default risk.

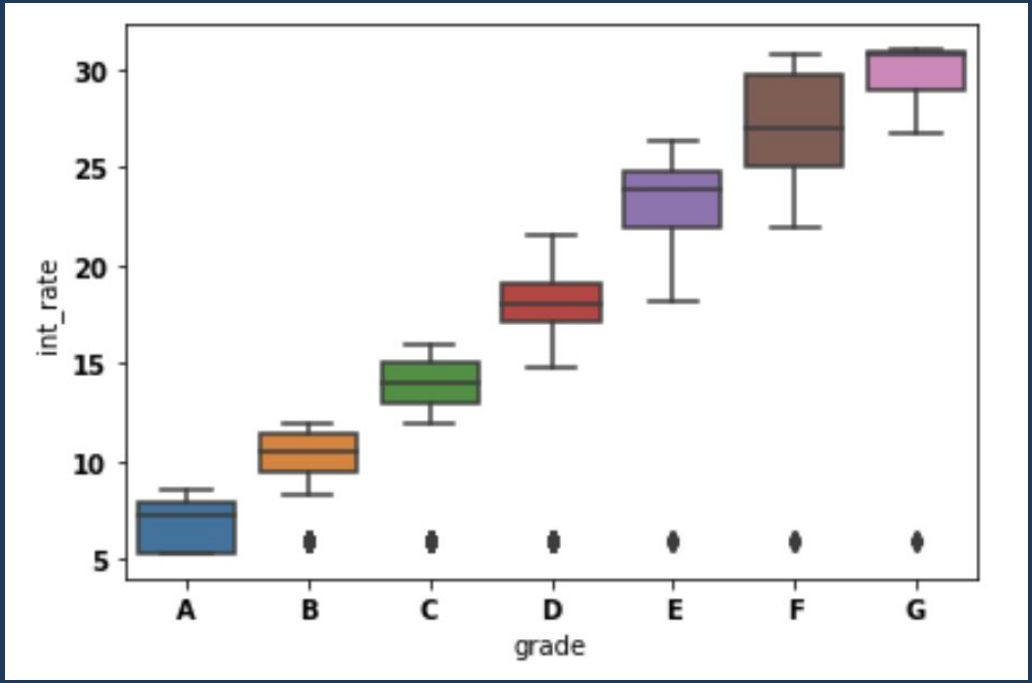


Income Level by Default Risk



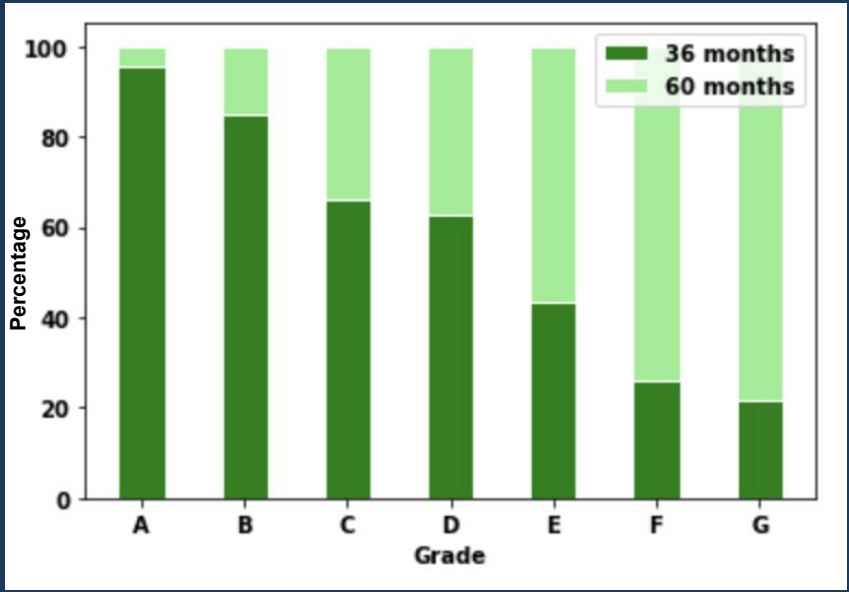
- A majority of the borrowers that tend to default on their loans have low income
- About half of the borrowers in our high risk class are mid_high and high income earners
- We observe a pretty even distribution between borrowers in our medium risk class

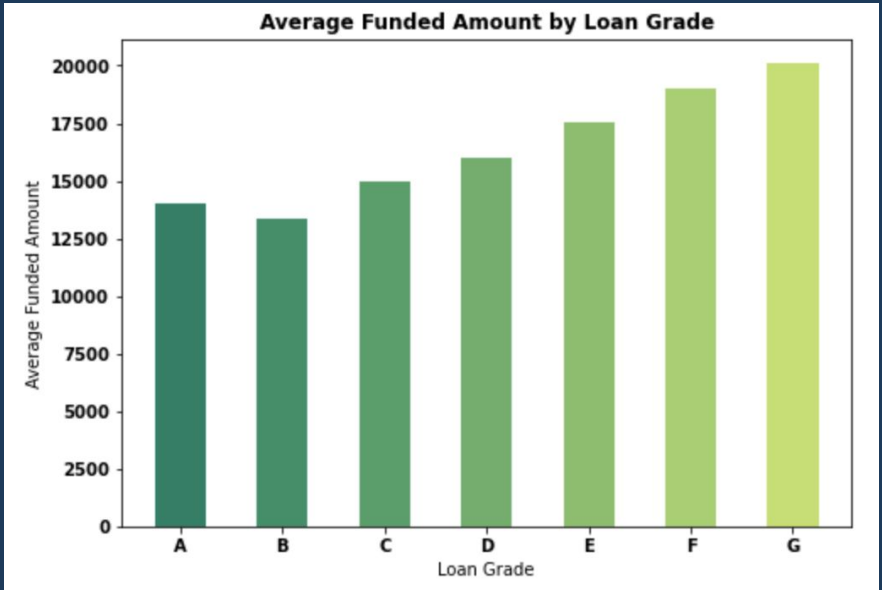
Loan Grade: Interest Rate



- People with lower grade has higher interest rate.
- There is a consistent group of outliers across loan grades.

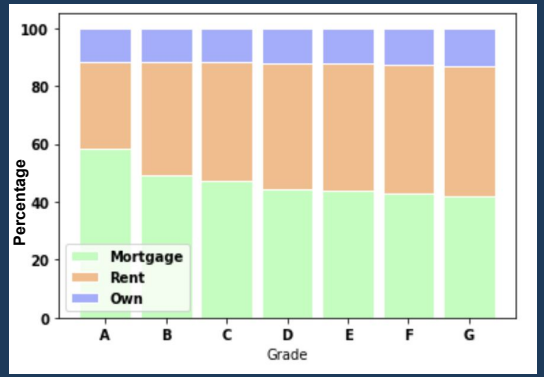
Loan Grade: Term & Average Funded Amount

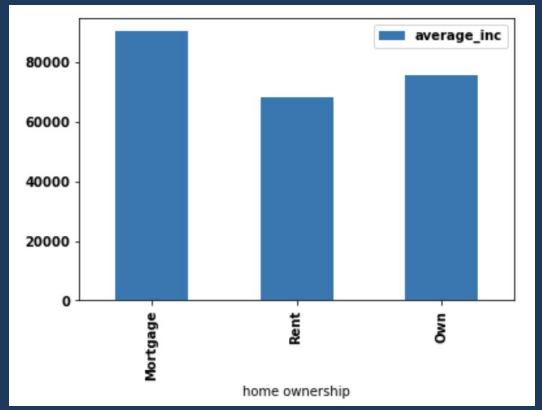


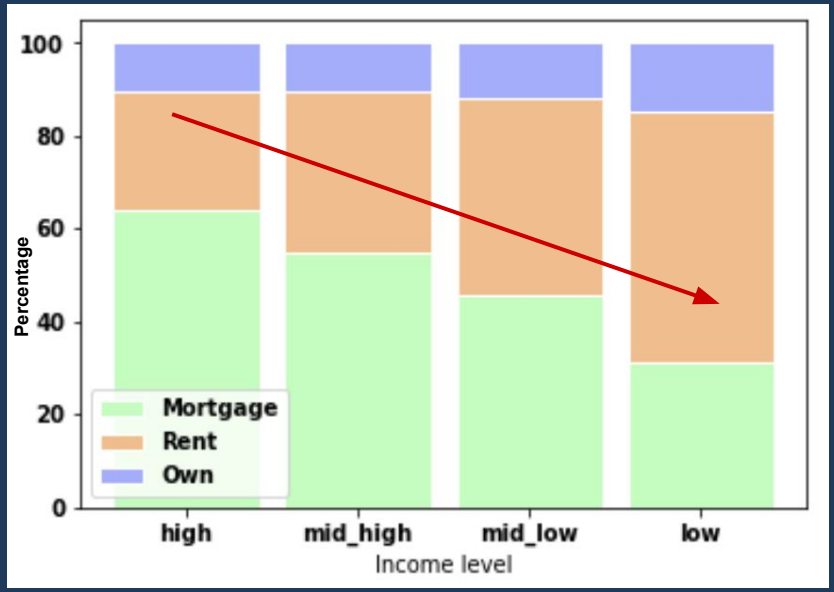


- Customers with higher loan grade tend to choose shorter term.
- Customers with lower loan grade have higher average funded amount.

Home Ownership



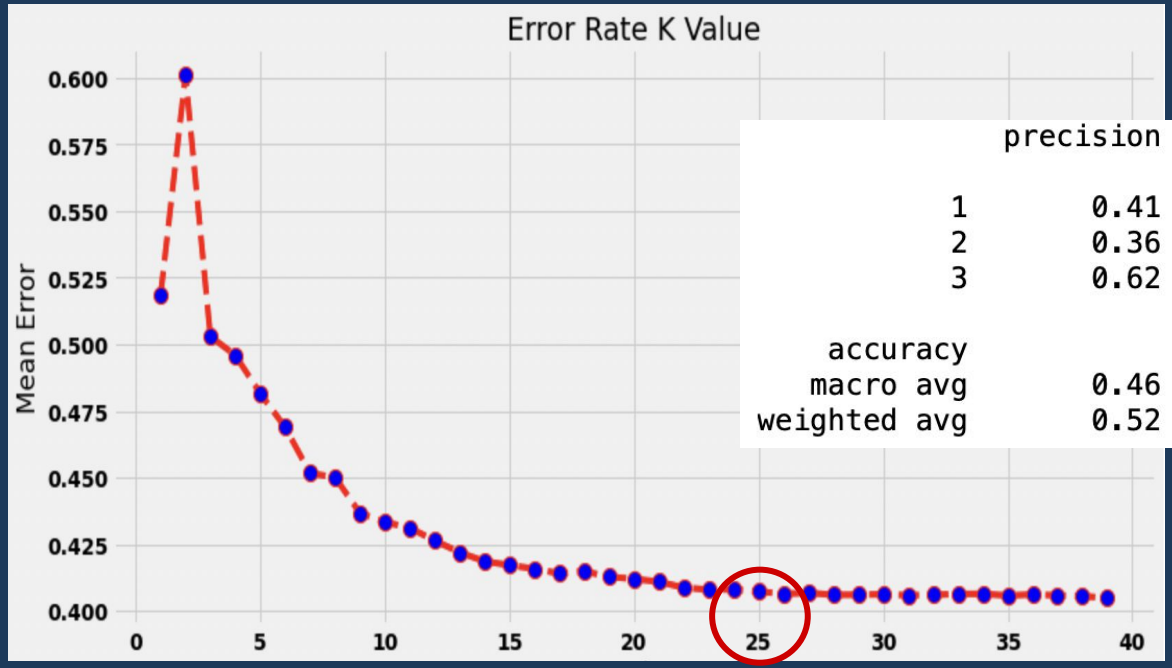




- 60% of high-income borrowers got a mortgage.
- Most low-income borrowers rent a house.

KNN

- Calculate error for K values between 1 and 40
- Choose K = 25 for K Nearest Neighbors

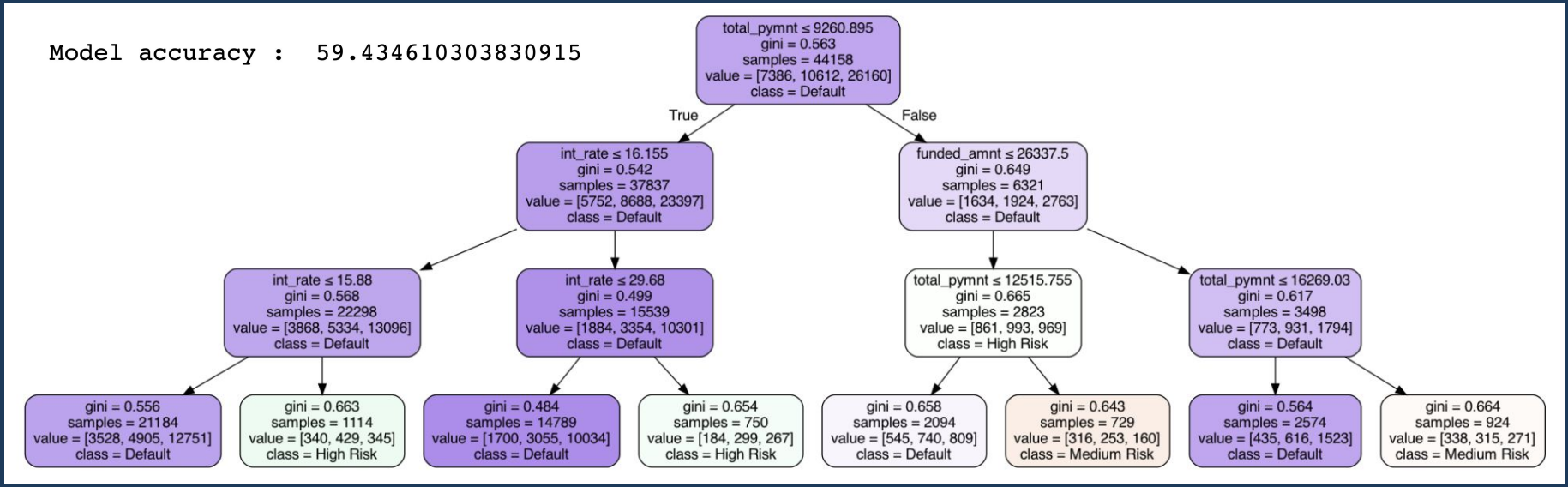


	precision	recall	f1-score	support
1	0.41	0.09	0.15	3121
2	0.36	0.11	0.17	4736
3	0.62	0.94	0.75	11068
accuracy			0.59	18925
macro avg	0.46	0.38	0.36	18925
weighted avg	0.52	0.59	0.50	18925

Decision Tree

- Use gini method
- Max length of 4
- Minimum 5 samples in leaf

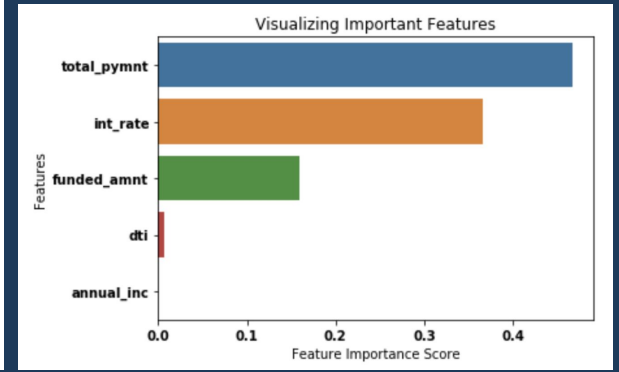
```
Model accuracy :  59.434610303830915
```



Decision Tree

	precision	recall	f1-score	support
1	0.43	0.11	0.18	3121
2	0.37	0.03	0.05	4736
3	0.61	0.97	0.75	11068
accuracy			0.59	18925
macro avg	0.47	0.37	0.33	18925
weighted avg	0.52	0.59	0.48	18925

```
total_pymnt      0.467342
int_rate         0.366342
funded_amnt      0.159532
dti              0.006785
annual_inc       0.000000
dtype: float64
```

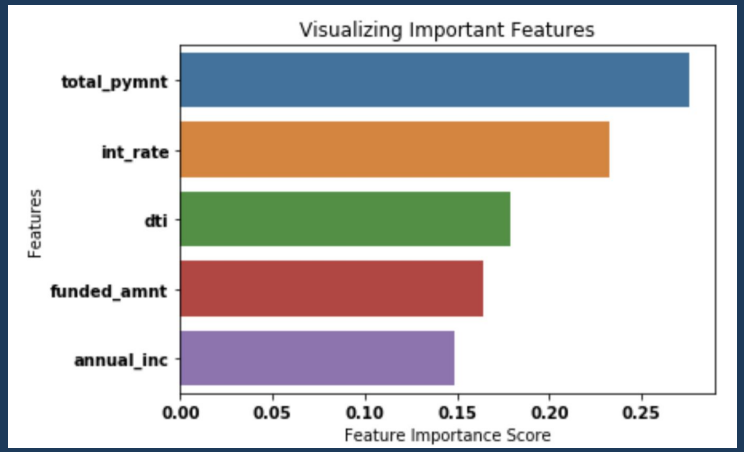


Random Forest

- n_estimators=100

```
Mean Absolute Error: 0.497623249669749
Mean Squared Error: 0.4047592708058124
Root Mean Squared Error: 0.6362069402370681
Accuracy: 70.6526992514311
```

```
total_pymnt     0.275983
int_rate        0.232288
dti             0.178589
funded_amnt     0.164160
annual_inc      0.148980
dtype: float64
```



Conclusion

- Learned a lot of about loans and the risk factors that lead to defaulting.
- 90% of the loans are fully paid back and in current status.
- Random forest was the best prediction model.
- All of the models accurately predicted default (class 3)
- Total payment and interest rate are the most important attributes to predict default risk.

Thank you!