Capstone Project

Lending Club Company Credit Risk

By: Cassandra Jones

NYC Data Science Academy

About Lending Club Company

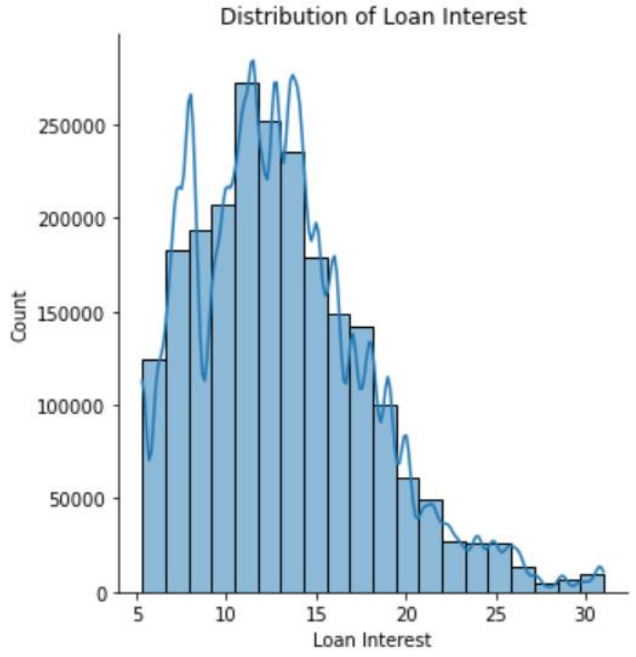
- Lending club is a financial-technology company that provides loans to client
- As a person working in a financial institute, I would like to know more detailed business trend of Lending Club by analyzing its historical data of loans from 2007 to 2018
- I will then use machine learning models on predicting the credit risk of the loans
- Dataset is from Kaggle with >2.2 million samples of loans issued by Lending Club

Data Cleaning

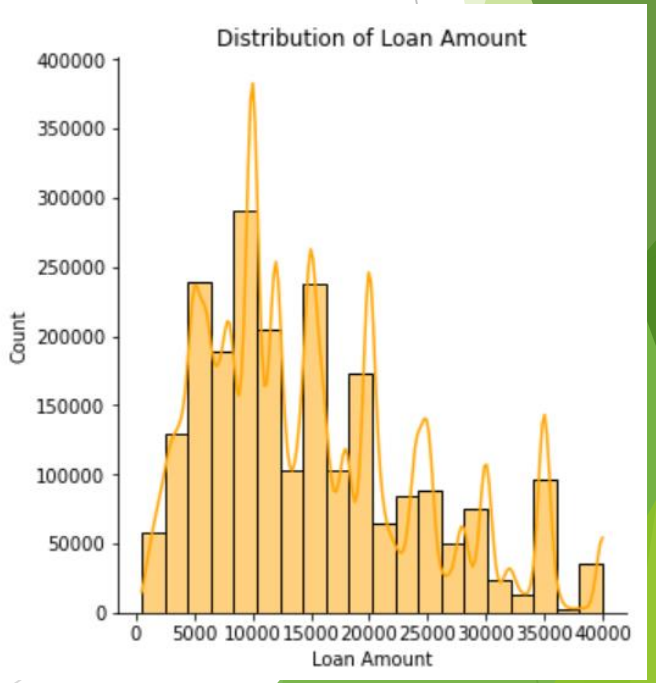
- There are 151 features from original dataset
- Drop features with >80% of missing value
- Drop features that do not affect the data analysis nor machine learning (i.e. zip code, loan ID, and etc.)
- Drop samples with >80% of missing value
- Numerical features: fill in median to the missing cells
- Categorical features: fill in mode to the missing cells

Data Analysis – Loan Interest & Amount

- Both loan interest and amount are right skewed

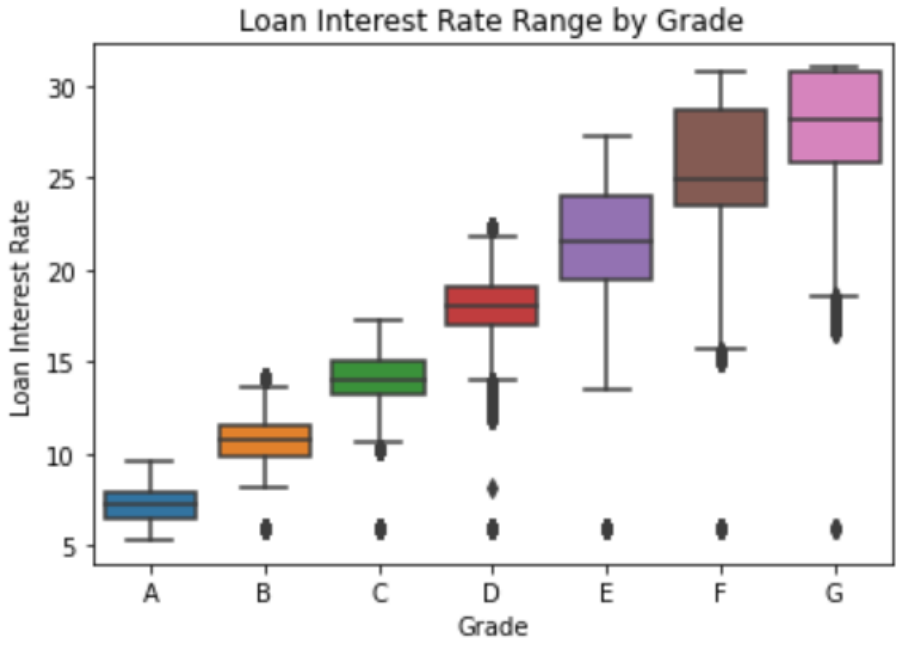


	int_rate	loan_amnt
count	2.260668e+06	2.260668e+06
mean	1.309283e+01	1.504693e+04
std	4.832138e+00	9.190245e+03
min	5.310000e+00	5.000000e+02
25%	9.490000e+00	8.000000e+03
50%	1.262000e+01	1.290000e+04
75%	1.599000e+01	2.000000e+04
max	3.099000e+01	4.000000e+04



Data Analysis – Grades

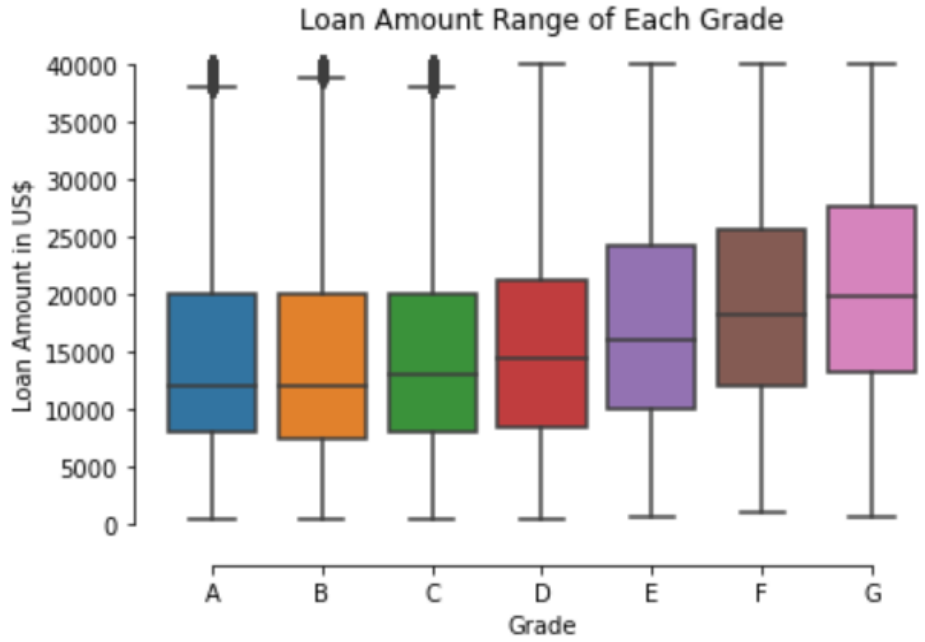
- Grade A has lowest range of interest rate while G has the highest range
- G is more profitable with higher risk.



grade	count	mean	std	min	25%	50%	75%	max
A	433027.0	7.084545	0.984465	5.31	6.46	7.24	7.89	9.63
B	663557.0	10.675806	1.238302	6.00	9.88	10.75	11.49	14.09
C	650053.0	14.143689	1.251283	6.00	13.22	13.99	15.02	17.27
D	324424.0	18.143067	1.676964	6.00	16.99	17.99	19.03	22.35
E	135639.0	21.829653	2.703925	6.00	19.52	21.48	23.99	27.27
F	41800.0	25.454091	2.928144	6.00	23.43	24.89	28.69	30.75
G	12168.0	28.074255	2.804587	6.00	25.89	28.18	30.79	30.99

Data Analysis – Grades

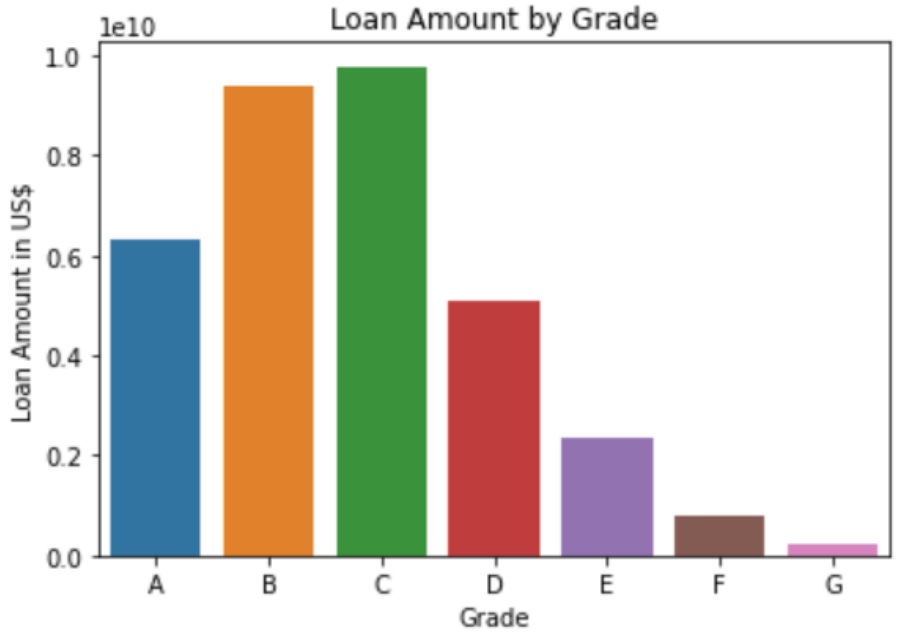
- F and G are having higher rates with less number of loans. However, the loan amounts are not less than other grades with even higher means.
- Does this mean E, F, and G



grade	count	mean	std	min	25%	50%	75%	max
A	433027.0	14603.343210	9107.975657	500.0	8000.00	12000.0	20000.00	40000.0
B	663557.0	14173.338199	8957.012601	500.0	7400.00	12000.0	20000.00	40000.0
C	650053.0	15038.083318	9203.950054	500.0	8000.00	13000.0	20000.00	40000.0
D	324424.0	15711.983007	9250.612823	500.0	8575.00	14400.0	21200.00	40000.0
E	135639.0	17453.078392	9363.276694	600.0	10000.00	16000.0	24175.00	40000.0
F	41800.0	19124.646531	9166.366254	1000.0	12000.00	18175.0	25600.00	40000.0
G	12168.0	20383.988741	8994.472986	600.0	13193.75	19800.0	27656.25	40000.0

Data Analysis – Grades

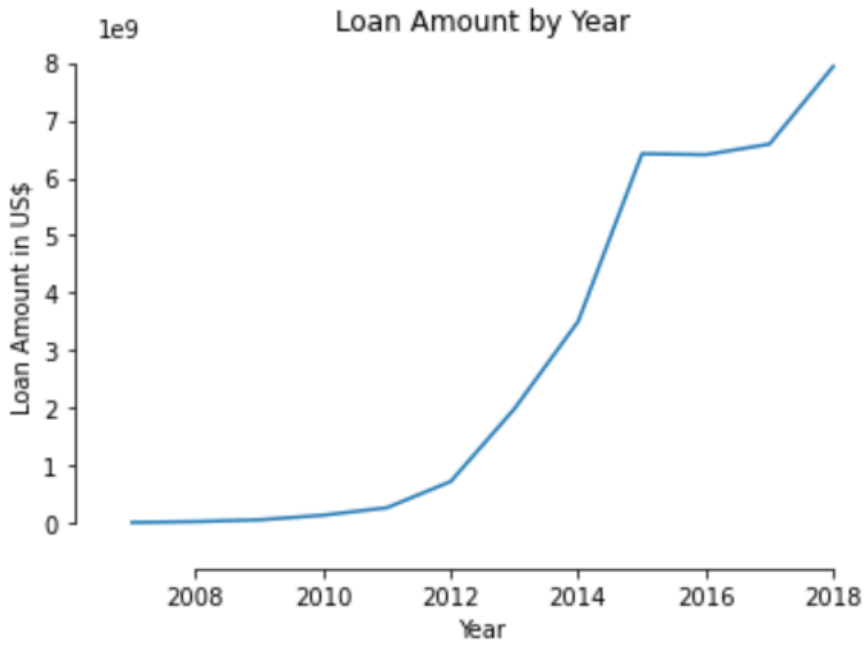
- Even though E, F, G seems more profitable and amounts of each transaction are larger from previous charts, company is having more business with B and C.

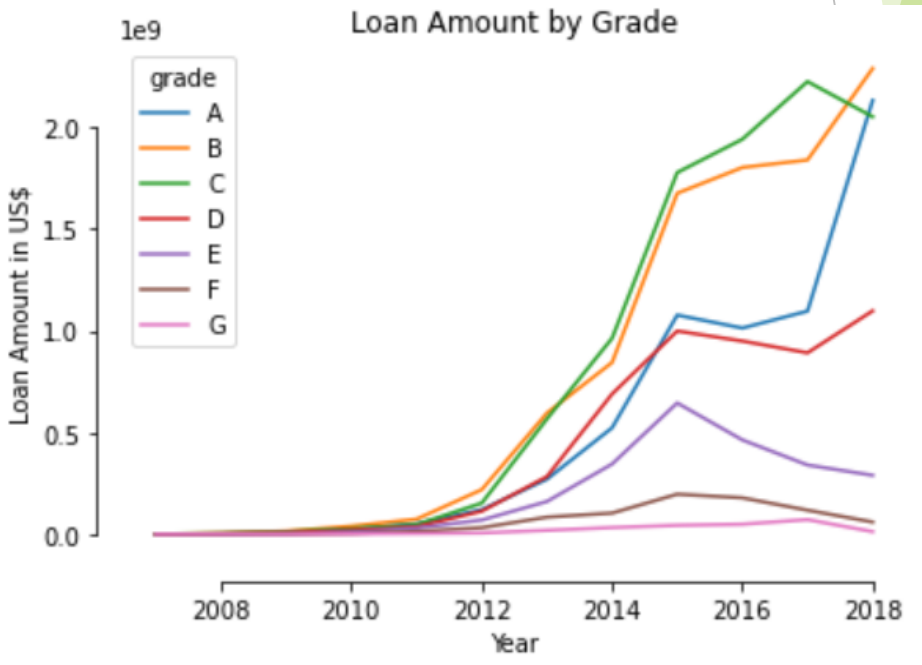


	loan_amnt	grade
0	6.323642e+09	A
1	9.404818e+09	B
2	9.775551e+09	C
3	5.097344e+09	D
4	2.367318e+09	E
5	7.994102e+08	F
6	2.480324e+08	G

Data Analysis – Trending

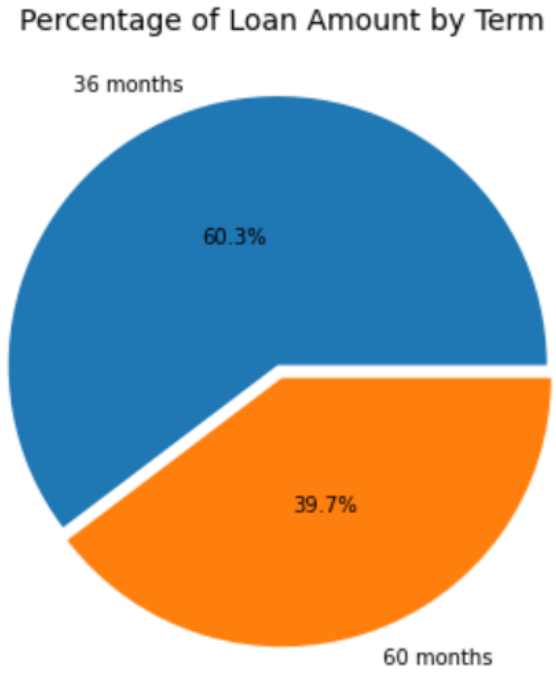
- Amount of loan grew steeply between 2012 – 2015 and after 2017 especially on grade A, B, C, and D, the loans with lower interest
- New York Times 2018: LendingClub Founder, Ousted in 2016, Settles Fraud Charges

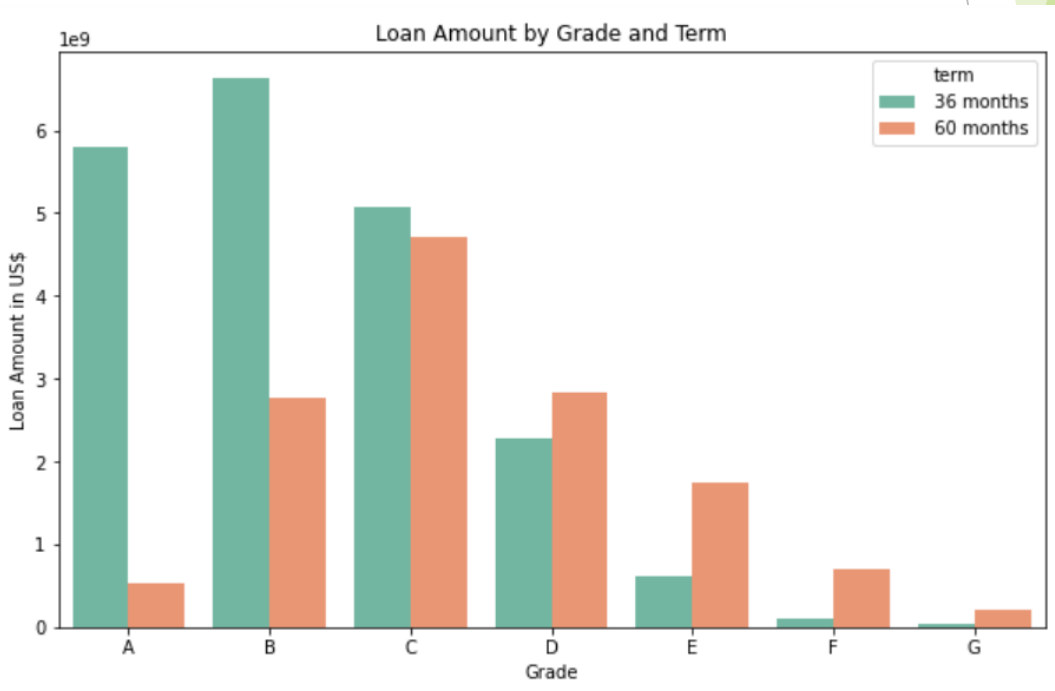




Data Analysis – Loan Terms

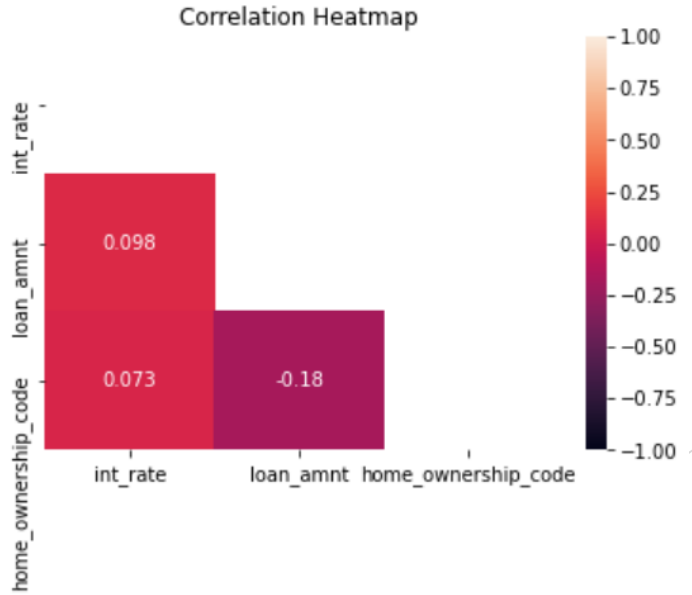
- 60.3% of loan amounts are under the tenor of 36 months with lower risk
- Shorter tenor lowers the interest rate





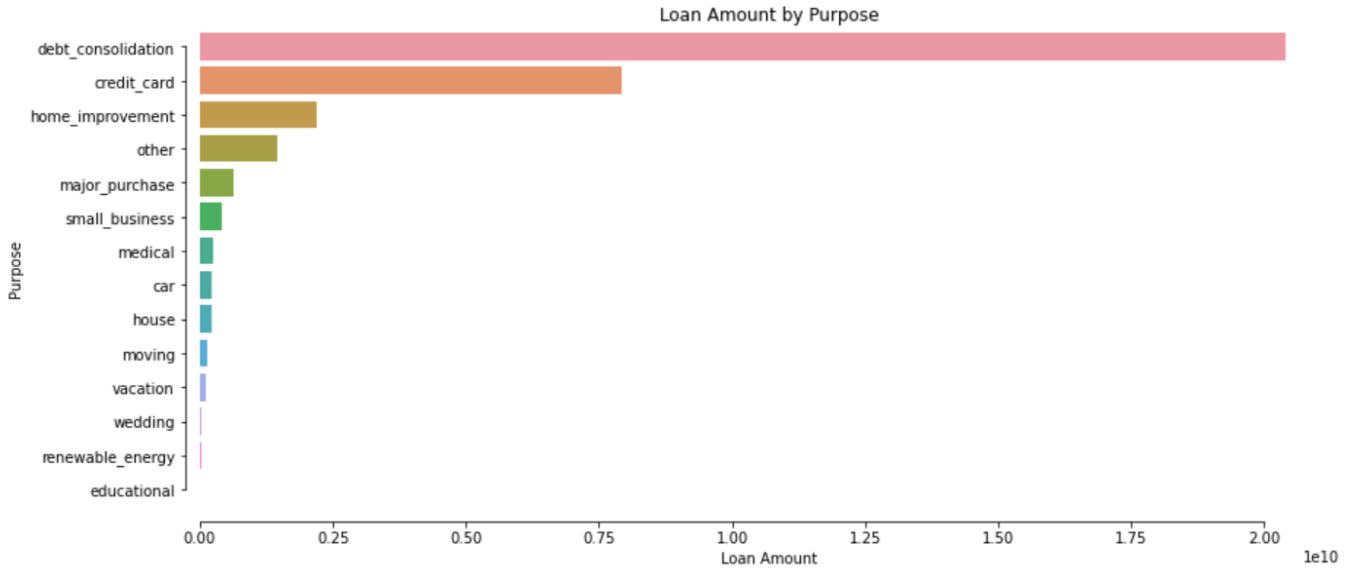
Data Analysis – Home Ownership

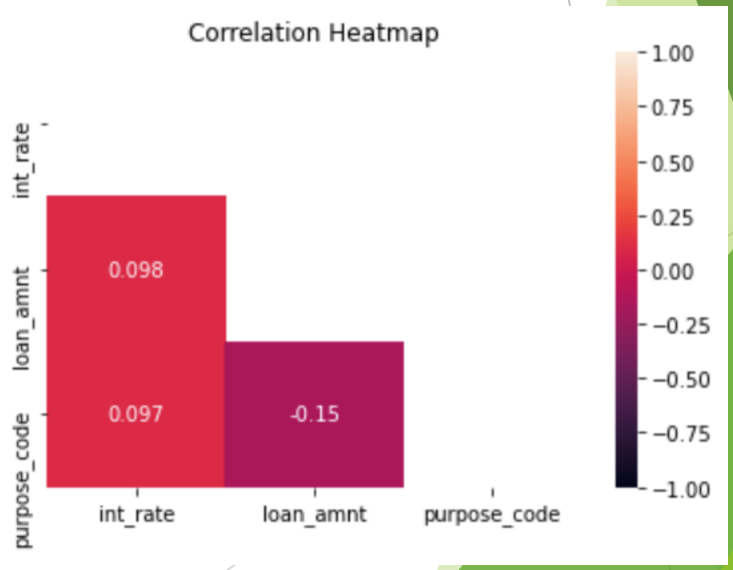
- Home ownership does not effect much on the loan interest nor amount



Data Analysis – Loan Purpose

- Most of the loans were used for debt consolidation and credit card
- Purpose of loans does not effect much on loan interest nor amount

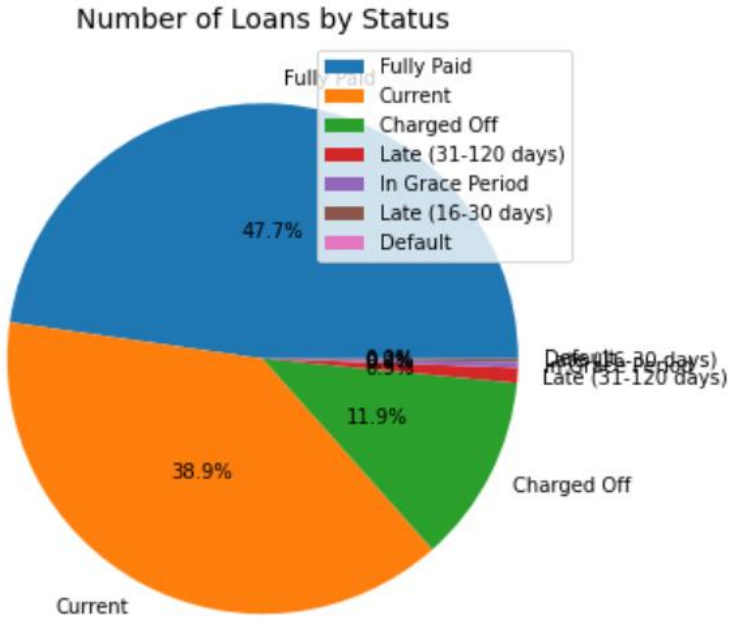


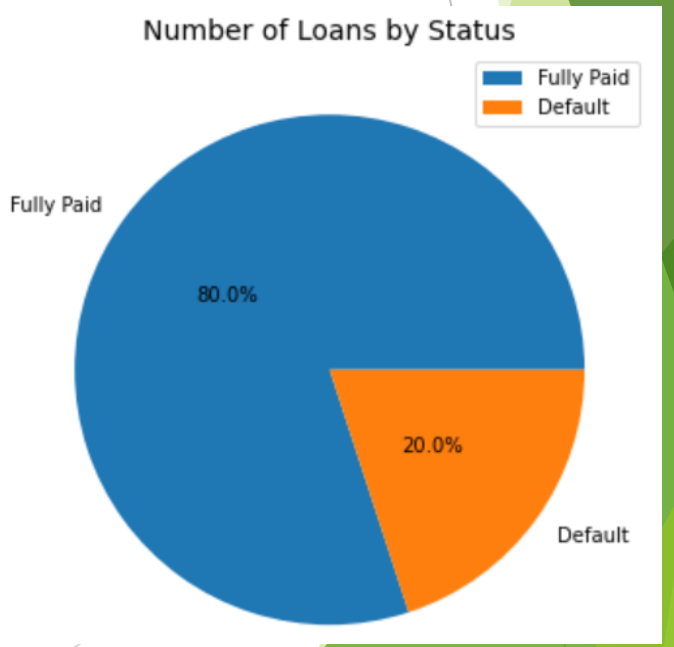


Data Analysis – Loan Status

- Charged off - company believes it will no longer collect as the borrower has become delinquent on payments
- Convert charged off to default as both are considered as bad debt and will be booked as reserve amount in accounting record

#_of_loans	loan_status
1078739	Fully Paid
878317	Current
269320	Charged Off
21467	Late (31-120 days)
8436	In Grace Period
4349	Late (16-30 days)
40	Default





Machine Learning - Preparation

- Select samples with loan status of “fully paid” and “default”
- Convert categorical features to numerical
- Remove features with multicolinearity for linear models

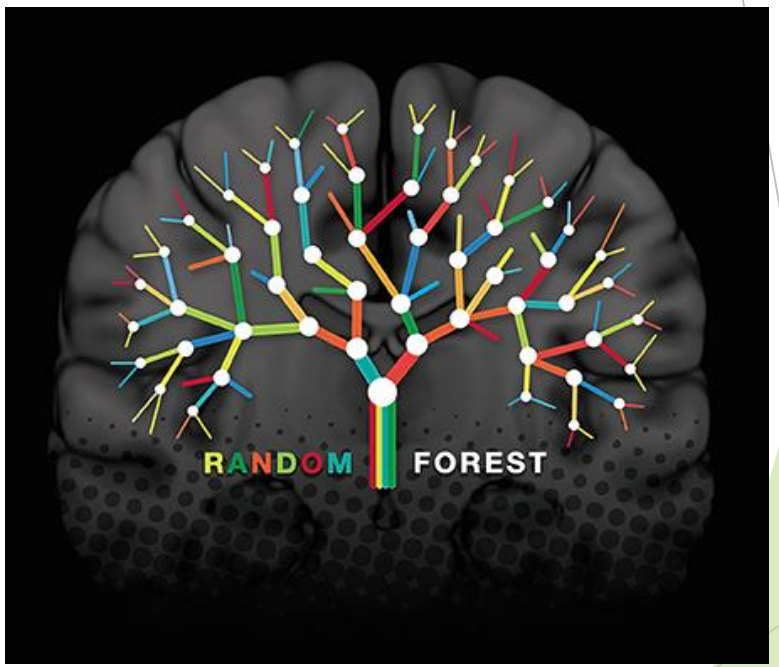
Machine Learning - Models

- Use imbalance-learn package to under sampling the train data size for random forest and gradient boosting by setting sampling strategy to 0.5
- Gradient boosting is taking too long to fit
- Train score of random forest is 0.9999965

model	train score	test score	recall score	ROC AUC score
Random Forest	1.0000	0.9969	0.9997	0.9998
Gradient Boosting	0.9962	0.9962	0.9998	0.9997
Logit	0.9310	0.9315	0.9287	0.9822
SVM	0.9144	0.7340	0.8888	0.5546
Naive Bayes	0.9073	0.9069	0.9610	0.9539

Conclusion

- Business of Lending Club is growing healthy with big increase of loan amount with lower risk (lower interest rate and shorter tenor)
- Among the closed transactions, 80% were having good credit
- Random Forest or Logit can be the models to predict the credit risk on outstanding transactions



The End

рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
mersi
kia ora
barka
welalin
tack
spas
dank je
misaotra
paldies
grazzi
mahalo
tapadh leat
faafetai lava
vinaka
blagodaram
matondo
kiitos
dankie
спасибі
thank you
gracias
хвала
asante
manana
dhanyavad
obrigada
tenki
bedankt
hvala
mauruuru
köszönöm
dankon
aciu
mochchakkeram
murakoze
enkosi
bayarlalaa
djiere dieuf
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
děkuji
sukriya
kop khun krap
taiku
obrigado
sagolun
arigatô
takk
dakujem
trugarez
mèsi
ありがとう
tanemirt
grazie
najis tuke
rahmet
terima kasih
diolch
dhanyavadagalu
shukriya
mercê
мерси
kam sah hamnida
xièxie
ευχαριστώ
merci
감사합니다