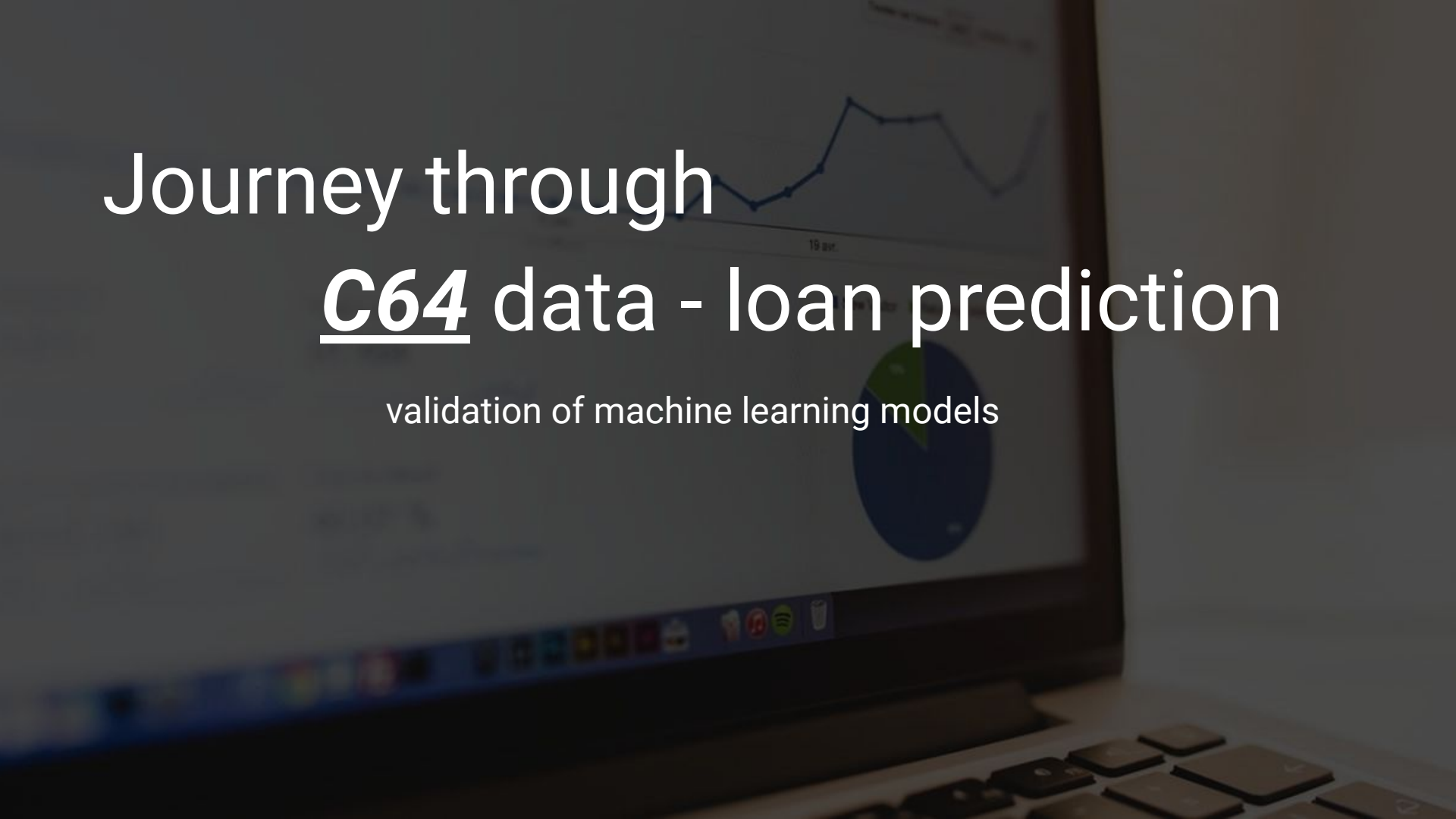
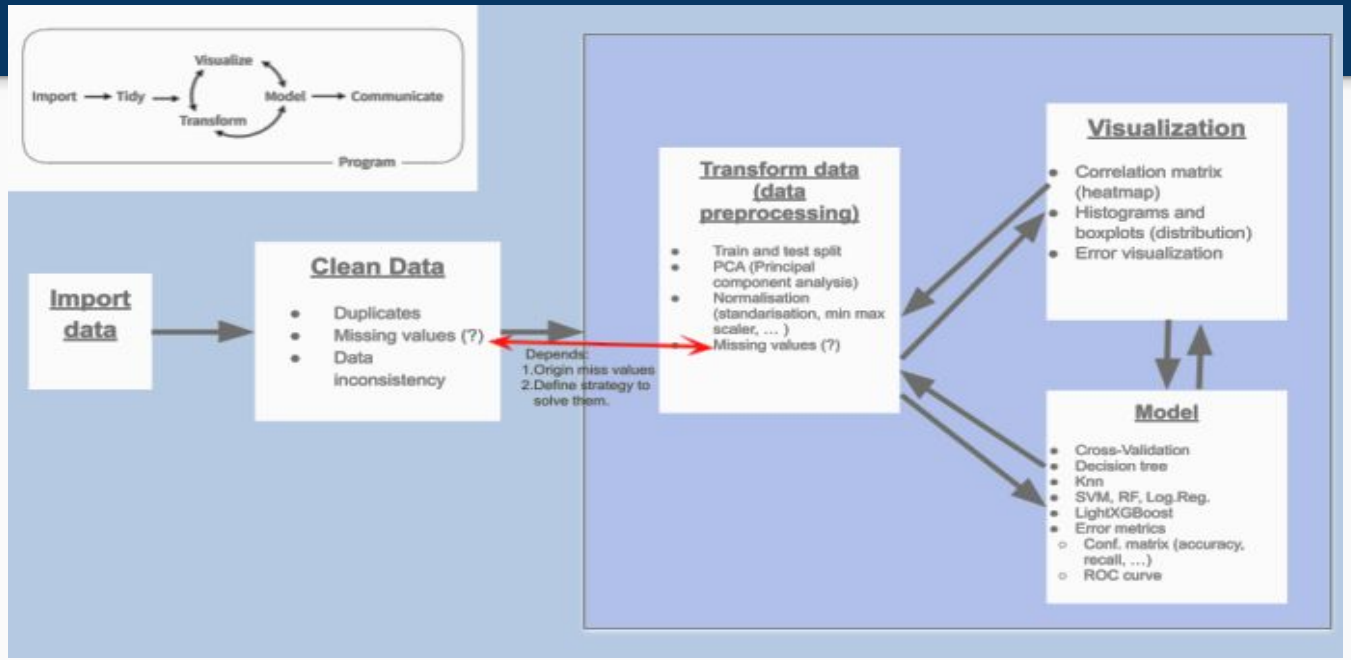


Journey through ***C64*** data - loan prediction

validation of machine learning models

The data pipeline (roadmap to success)



C64 dataset

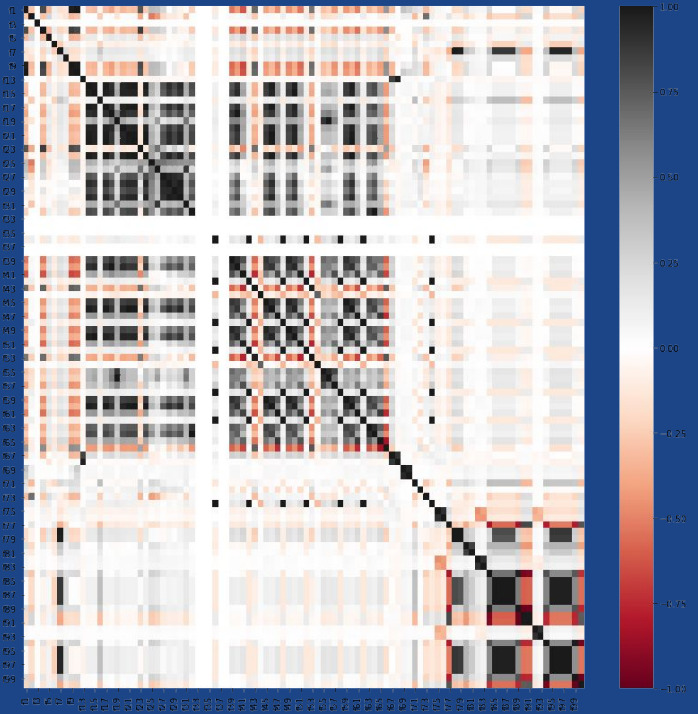
id	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	...	f93	f94	f95	f96	f97	f98	f99	f100	loss	target
1	126	10	0.686842	1100	3	13699	7201.0	4949.0	126.75	126.03	...	0.01	0.12	851.0	851.0	851.0	851.0	0.32	1.34	0	paid
2	121	10	0.782776	1100	3	84645	240.0	1625.0	123.52	121.35	...	0.09	0.08	20.0	20.0	20.0	20.0	0.28	1.43	0	paid
3	126	10	0.500080	1100	3	83607	1800.0	1527.0	127.76	126.49	...	0.00	0.07	124.0	124.0	124.0	124.0	0.25	1.52	0	paid
4	134	10	0.439874	1100	3	82642	7542.0	1730.0	132.94	133.58	...	0.00	0.12	903.0	903.0	903.0	903.0	0.32	1.35	0	paid
5	109	9	0.502749	2900	4	79124	89.0	491.0	122.72	112.77	...	0.90	0.06	5.0	5.0	5.0	5.0	0.23	1.56	0	paid
6	126	9	0.691954	2900	4	14448	1514.0	4176.0	127.74	126.23	...	0.00	0.09	131.0	131.0	131.0	131.0	0.28	1.43	0	paid
7	121	9	0.985674	2900	4	13026	4565.0	263.0	126.36	122.09	...	0.00	0.09	399.0	399.0	399.0	399.0	0.28	1.44	0	paid
8	128	9	0.385778	2900	4	79244	6597.0	3592.0	127.19	127.89	...	0.09	0.13	837.0	837.0	837.0	837.0	0.33	1.32	1	not paid
9	126	9	0.745471	2900	4	78920	3058.0	112.0	123.89	125.53	...	0.00	0.08	253.0	253.0	253.0	253.0	0.28	1.43	0	paid
10	127	9	0.580561	2900	4	83442	684.0	1141.0	127.00	127.39	...	0.00	0.12	82.0	82.0	82.0	82.0	0.32	1.35	0	paid
11	115	9	0.611158	1300	3	6901	685.0	2437.0	117.25	115.03	...	0.90	0.12	83.0	83.0	83.0	83.0	0.33	1.31	0	paid
12	120	9	0.801255	1300	3	13026	4566.0	982.0	118.14	119.56	...	0.00	0.09	399.0	399.0	399.0	399.0	0.28	1.44	0	paid
13	130	9	0.574090	1300	3	8563	5264.0	3566.0	127.10	129.17	...	0.00	0.12	656.0	656.0	656.0	656.0	0.33	1.32	0	paid
14	119	9	0.445187	1300	3	2456	8236.0	983.0	122.79	119.79	...	0.00	0.08	650.0	650.0	650.0	650.0	0.27	1.46	0	paid
15	116	9	0.092193	1300	3	83726	3059.0	1731.0	121.90	116.94	...	0.00	0.08	253.0	253.0	253.0	253.0	0.28	1.43	0	paid

Cleaning process

id	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	...	f93	f94	f95	f96	f97	f98	f99	f100	loss	target
1	126	10	0.686842	1100	3	13699	7201.0	4949.0	126.75	126.03	...	0.01	0.12	851.0	851.0	851.0	851.0	0.32	1.34	0	paid
2	121	10	0.782776	1100	3	84645	240.0	1625.0	123.52	121.35	...	0.09	0.08	20.0	20.0	20.0	20.0	0.28	1.43	0	paid
3	126	10	0.500080	1100	3	83607	1800.0	1527.0	127.76	126.49	...	0.00	0.07	124.0	124.0	124.0	124.0	0.25	1.52	0	paid
4	134	10	0.439874	1100	3	82642	7542.0	1730.0	132.94	133.58	...	0.00	0.12	903.0	903.0	903.0	903.0	0.32	1.35	0	paid
5	109	9	0.502749	2900	4	79124	89.0	491.0	122.72	112.77	...	0.90	0.06	5.0	5.0	5.0	5.0	0.23	1.56	0	paid
6	126	9	0.691954	2900	4	14448	1514.0	4176.0	127.74	126.23	...	0.00	0.09	131.0	131.0	131.0	131.0	0.28	1.43	0	paid
7	121	9	0.985674	2900	4	13026	4565.0	263.0	126.36	122.09	...	0.00	0.09	399.0	399.0	399.0	399.0	0.28	1.44	0	paid
8	128	9	0.385778	2900	4	79244	6597.0	3592.0	127.19	127.89	...	0.09	0.13	837.0	837.0	837.0	837.0	0.33	1.32	1	not paid
9	126	9	0.745471	2900	4	78920	3058.0	112.0	123.89	125.53	...	0.00	0.08	253.0	253.0	253.0	253.0	0.28	1.43	0	paid
10	127	9	0.580561	2900	4	83442	684.0	1141.0	127.00	127.39	...	0.00	0.12	82.0	82.0	82.0	82.0	0.32	1.35	0	paid
11	115	9	0.611158	1300	3	6901	685.0	2437.0	117.25	115.03	...	0.90	0.12	83.0	83.0	83.0	83.0	0.33	1.31	0	paid
12	120	9	0.801255	1300	3	13026	4566.0	982.0	118.14	119.56	...	0.00	0.09	399.0	399.0	399.0	399.0	0.28	1.44	0	paid
13	130	9	0.574090	1300	3	8563	5264.0	3566.0	127.10	129.17	...	0.00	0.12	656.0	656.0	656.0	656.0	0.33	1.32	0	paid
14	119	9	0.445187	1300	3	2456	8236.0	983.0	122.79	119.79	...	0.00	0.08	650.0	650.0	650.0	650.0	0.27	1.46	0	paid
15	116	9	0.092193	1300	3	83726	3059.0	1731.0	121.90	116.94	...	0.00	0.08	253.0	253.0	253.0	253.0	0.28	1.43	0	paid

- ~ 8,6% missing values
- no duplicates
- shape before cleaning : (80000, 99)
- shape after cleaning : (68708, 99)

Data preprocessing



- Feature creation
  + creating col. target based on col. loss => paid / not paid loan

paid	62262
not paid	6446

- Feature selection
  + exclude col. with near zero variance from 100 col. => 93 col.
  + normalise datasets

- Balance the dataset
  + split into two manageable datasets

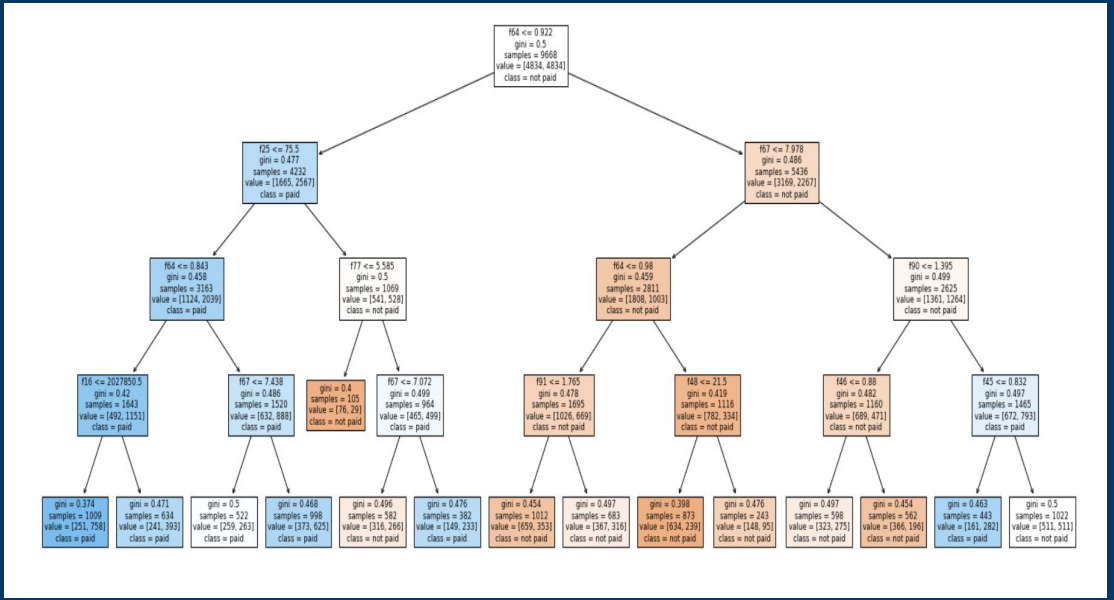
The size of the train_paid_data is the same as for the train_not_paid_data: 4834

& create balanced train and test sets

Decision Tree

- train set accuracy_score 0.6286719073231278
- test set accuracy_score 0.5809112466124661
- confusion matrix on test set

	not paid	paid
not paid	984	628
paid	24115	33313



KNN

First Iteration

- train set accuracy_score
  0.630947455523376
- test set accuracy_score
  0.5158705962059621
- confusion matrix on test set

	not paid	paid
not paid	882	730
paid	27853	29575

Second Iteration

- train set accuracy_score
  0.6634257343814646
- test set accuracy_score
  0.5415650406504066
- confusion matrix on test set

	not paid	paid
not paid	970	642
paid	26424	31004

Logistic regression

- train set accuracy_score
  0.6291890773686388

- test set accuracy_score
  0.6006944444444444

- confusion matrix on test set

	not paid	paid
not paid	1025	587
paid	22988	34440

Random forest

- train set accuracy_score
  0.6152254861398427
- test set accuracy_score
  0.5531673441734417
- confusion matrix on test set

	not paid	paid
not paid	1052	560
paid	25821	31607

Conclusion

Decision Tree

	not paid	paid
not paid	984	628
paid	24115	33313

KNN (2.Iteration)

	not paid	paid
not paid	970	642
paid	26424	31004

Logistic regression

	not paid	paid
not paid	1025	587
paid	22988	34440

Random forest

	not paid	paid
not paid	1052	560
paid	25821	31607

train | test

0.628| 0.580

0.663 | 0.541

0.629 | 0.600

0.615 | 0.553