HOME CREDIT DEFAULT RISK

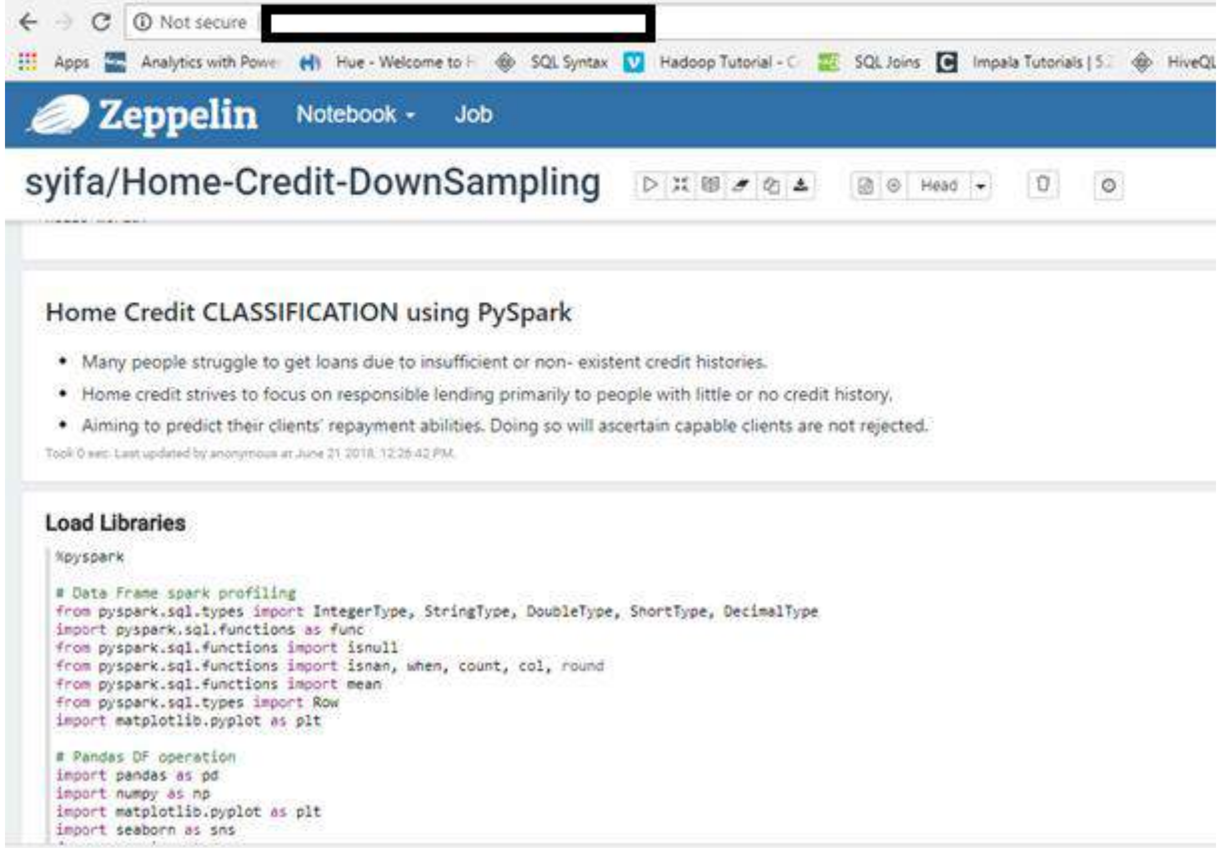
USING SPARK AND ZEPPELIN

HOME CREDIT

OVERVIEW

- Many people struggle to get loans due to **insufficient** or **non- existent credit histories**.
- Home credit **strives** to focus on responsible lending primarily to people with little or no credit history.
- Aiming **to predict** their clients' repayment abilities. Doing so will ascertain capable clients are not rejected.

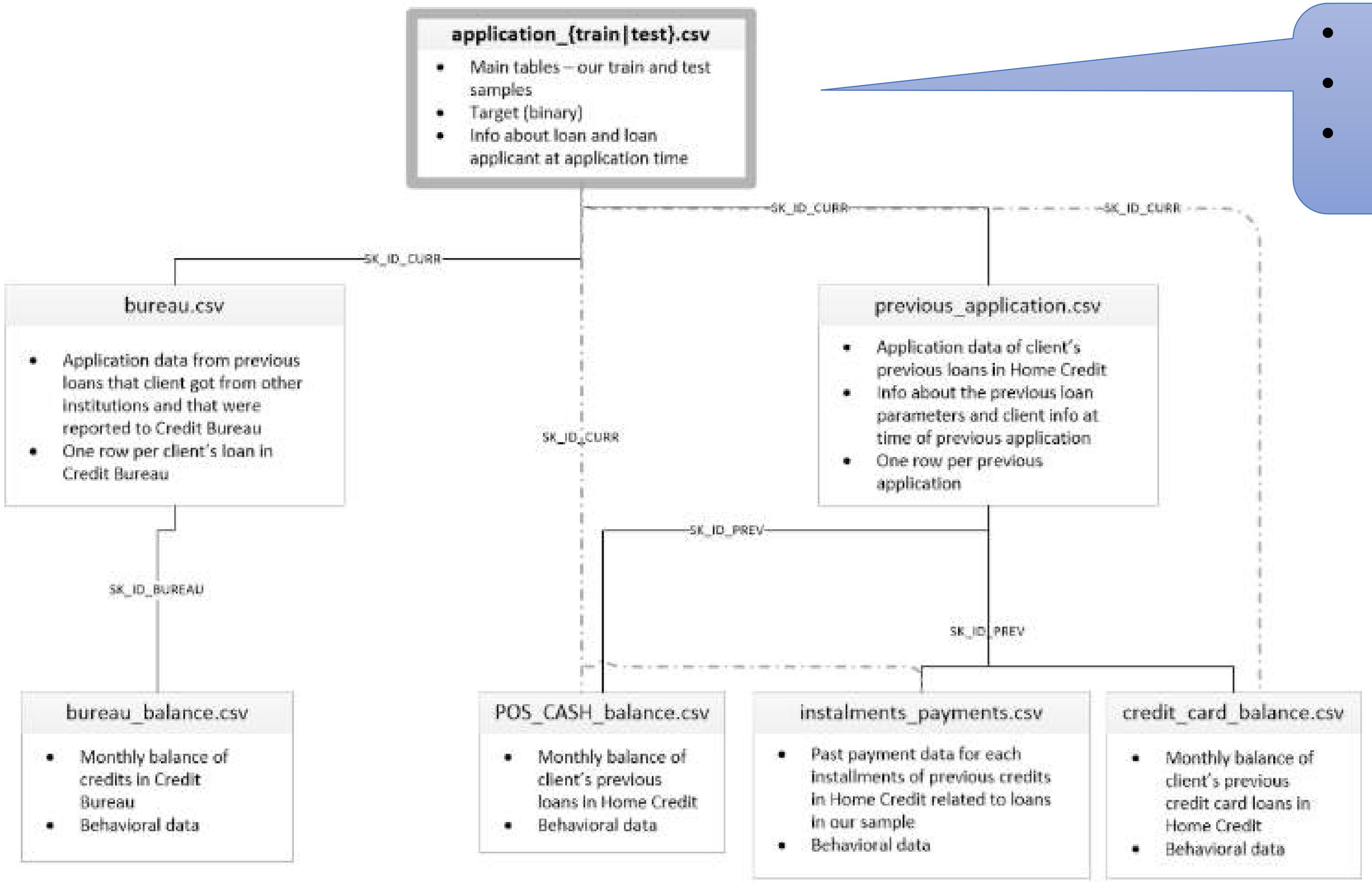
ZEPPELIN & SPARK



- Interactive web-based notebooks.
- Data ingestion, data exploration, visualization and sharing data.

- Analytic engine for big data processing.
- Easy of use: Java, Scala, Python, R, and SQL.
- Speed, run workloads 100x faster.

DATA



- Focus here.
- To build up understanding.
- Avoid diving all and getting lost.

HOME CREDIT

GLIMPSE of DATA

Let's focus at data train (hc_train)
Rename Columns
Selected Columns:

```
|-- SK_ID_CURR: integer (nullable = true)
|-- label: integer (nullable = true)
|-- CONTRACT_TYPE: string (nullable = true)
|-- GENDER: string (nullable = true)
|-- FLAG_OWN_CAR: string (nullable = true)
|-- CNT_CHILDREN: integer (nullable = true)
|-- AMT_INCOME_TOTAL: double (nullable = true)
|-- AMT_CREDIT: double (nullable = true)
|-- AMT_ANNUITY: double (nullable = true)
|-- INCOME_TYPE: string (nullable = true)
|-- EDUCATION: string (nullable = true)
|-- MARRIAGE: string (nullable = true)
|-- HOUSING_TYPE: string (nullable = true)
|-- DAYS_BIRTH: integer (nullable = true)
|-- OCCUPATION: string (nullable = false)
|-- CNT_FAM_MEMBERS: double (nullable = false)
|-- EXT_SOURCE_1: double (nullable = true)
|-- EXT_SOURCE_2: double (nullable = true)
|-- EXT_SOURCE_3: double (nullable = true)
```

Shape:

- hc_train: **307511, 122**
- hc_test: **48744, 121**

Categorical Variables: 8
Numerical Variables: 9

GLIMPSE of DATA

SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	AMT_INCOME_TOTAL	AMT_CREDIT	NAME_INCOME_TYPE	E	CNT_FAM_MEMBERS
100002	1	Cash loans	M	N	202500.0	406597.5	Working	s	1.0
100003	0	Cash loans	F	N	270000.0	1293502.5	State servant	f	2.0
100004	0	Revolving loans	M	Y	67500.0	135000.0	Working	s	1.0
100006	0	Cash loans	F	N	135000.0	312682.5	Working	s	2.0
100007	0	Cash loans	M	N	121500.0	513000.0	Working	f	1.0

Negative values means before the day of application.

Show 5 observations from data train (selected columns).

NAME_EDUCATION_TYPE	NAME_FAMILY_STATUS	NAME_HOUSING_TYPE	DAYS_BIRTH	OCCUPATION_TY
Secondary / secon...	Single / not married	House / apartment	-9461	Labore
Higher education	Married	House / apartment	-16765	Core sta
Secondary / secon...	Single / not married	House / apartment	-19046	Labore
Secondary / secon...	Civil marriage	House / apartment	-19005	Labore
Secondary / secon...	Single / not married	House / apartment	-19932	Core sta

MISSING VALUE

Variables which have missing value:

OCCUPATION	COUNT
true	96391
false	211120

EXT_SOURCE_1	COUNT
true	173378
false	134133

CNT_FAM_MEMBERS	COUNT
true	2
false	307509

EXT_SOURCE_2	COUNT
true	660
false	306851

AMT_ANNUITY	COUNT
true	12
false	307499

EXT_SOURCE_2	COUNT
true	60965
false	246546

FILL MISSING VALUE

Method to fill missing value:

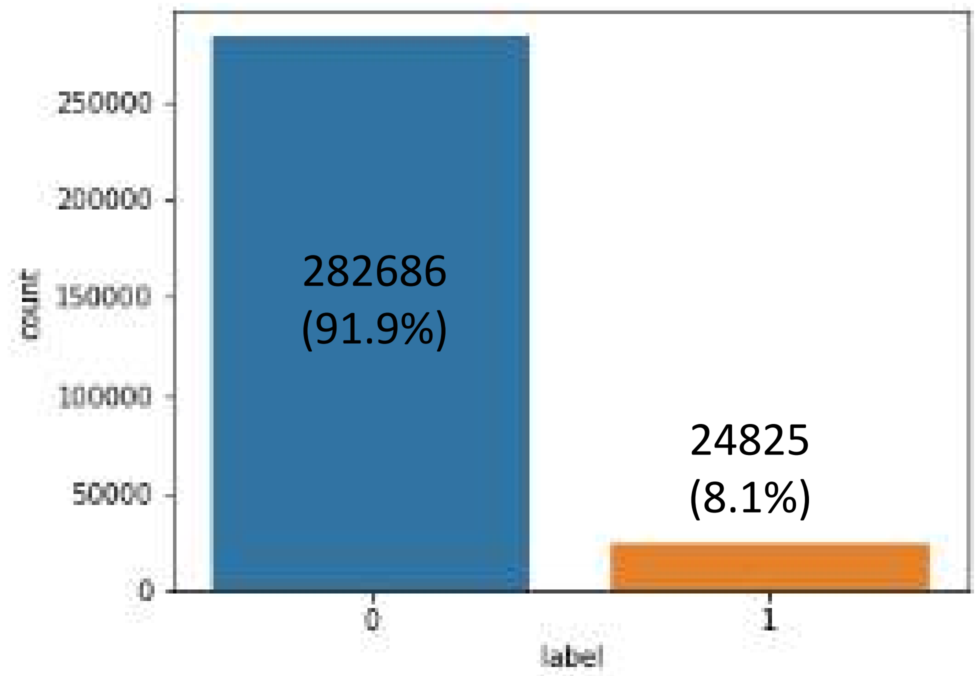
- For categorical variables → use mode (most frequent of category) to fill missing value
- For numerical variables → use mean or average to fill missing value, forward fill and back fill.

Fill na with :"Laborers" → OCCUPATION

Fill na with : average → CNT_FAM_MEMBERS, AMT_ANNUITY, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3

VISUALIZATION

Data Exploration



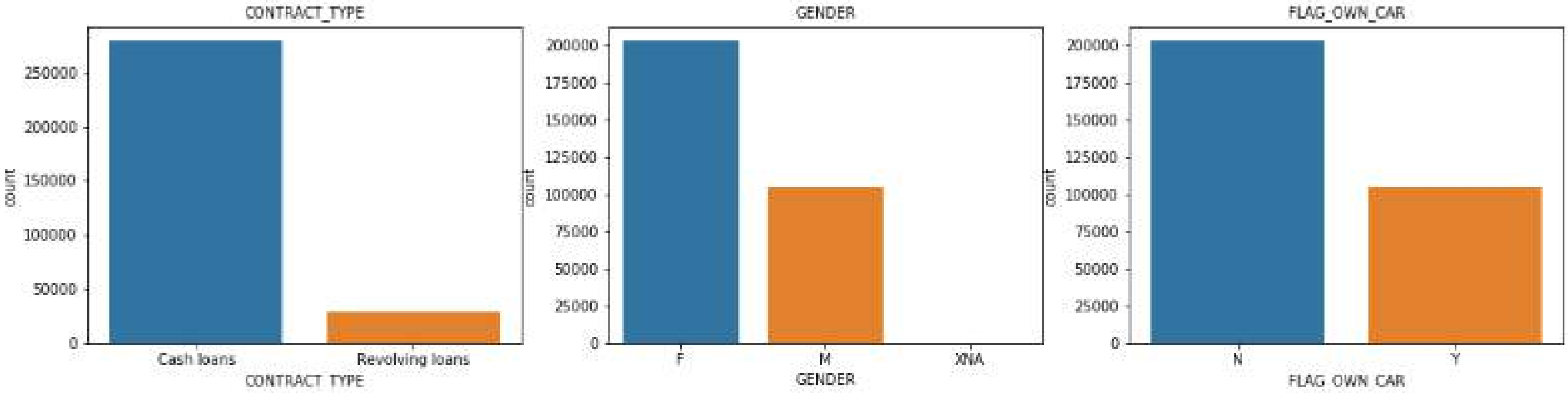
Imbalance classification.

Imbalance classification is a condition where the difference number of observations between one class with other class is **huge**. There are some method to handle imbalance data:

- Down sampling
- Over sampling

VISUALIZATION

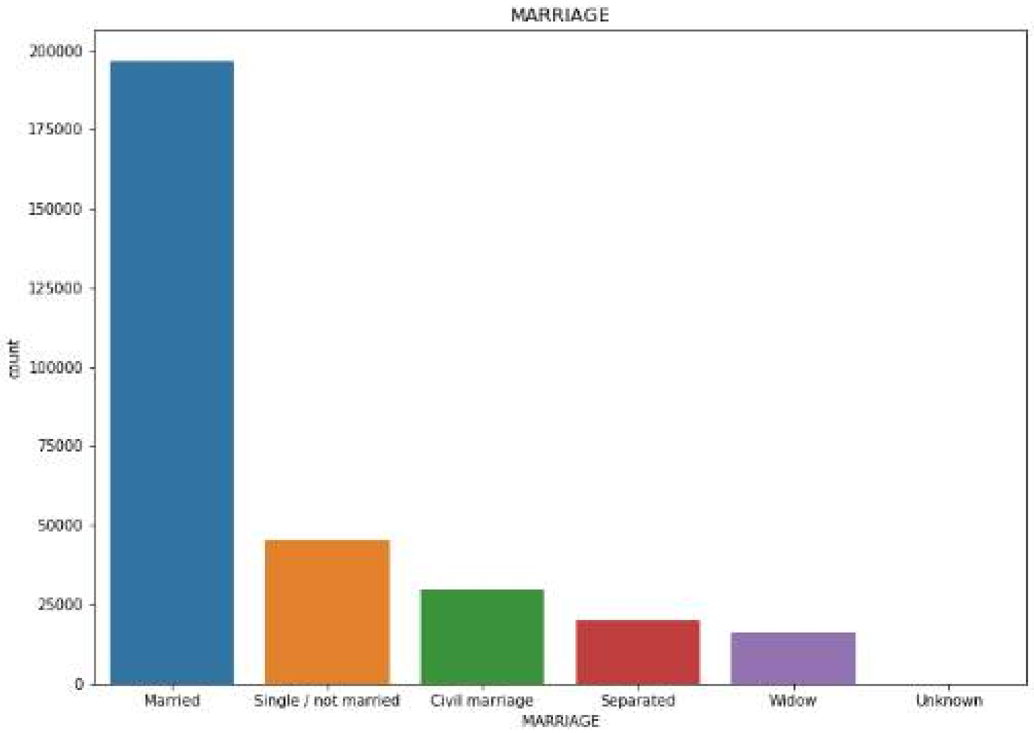
Data Exploration: Contract type, gender and flag own car.

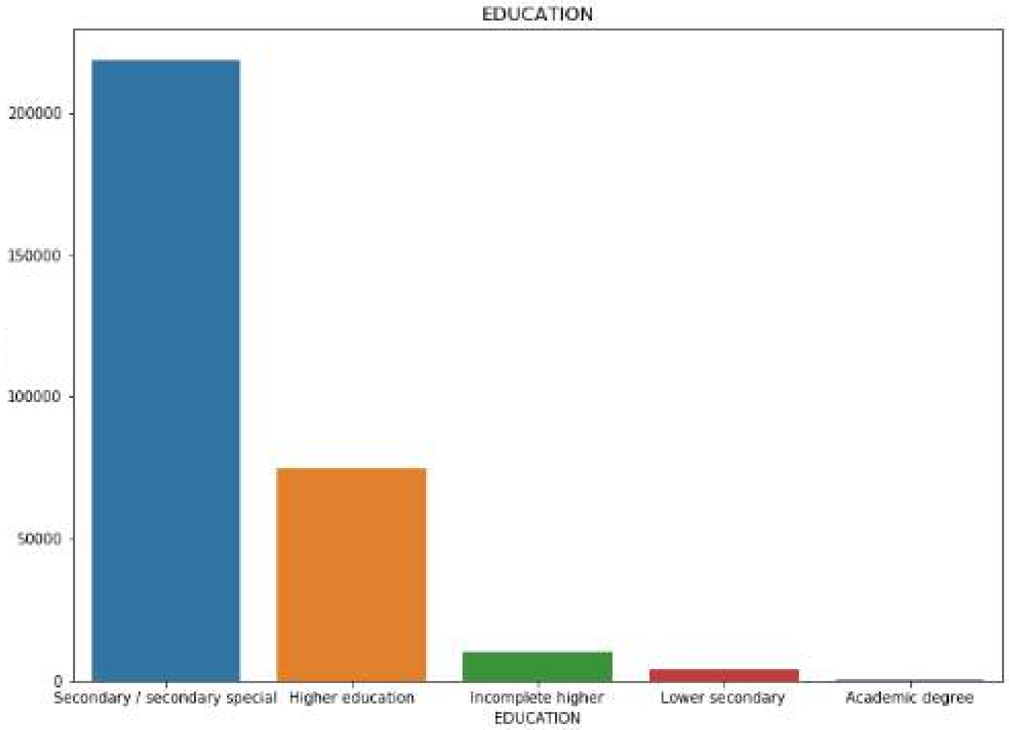


Categories with lowest quantities can be combined with other category.

VISUALIZATION

Data Exploration: Marriage and education.

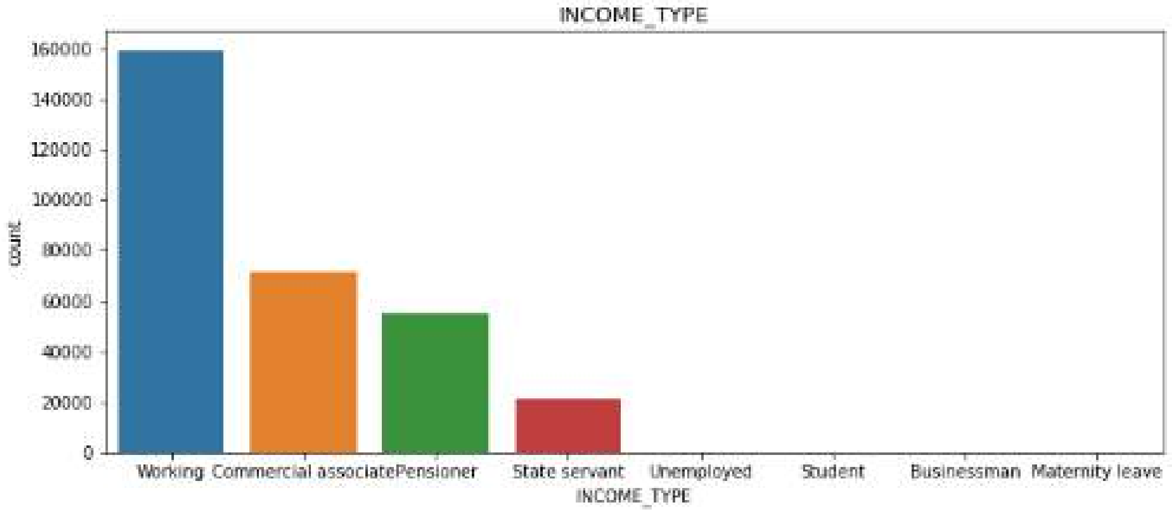


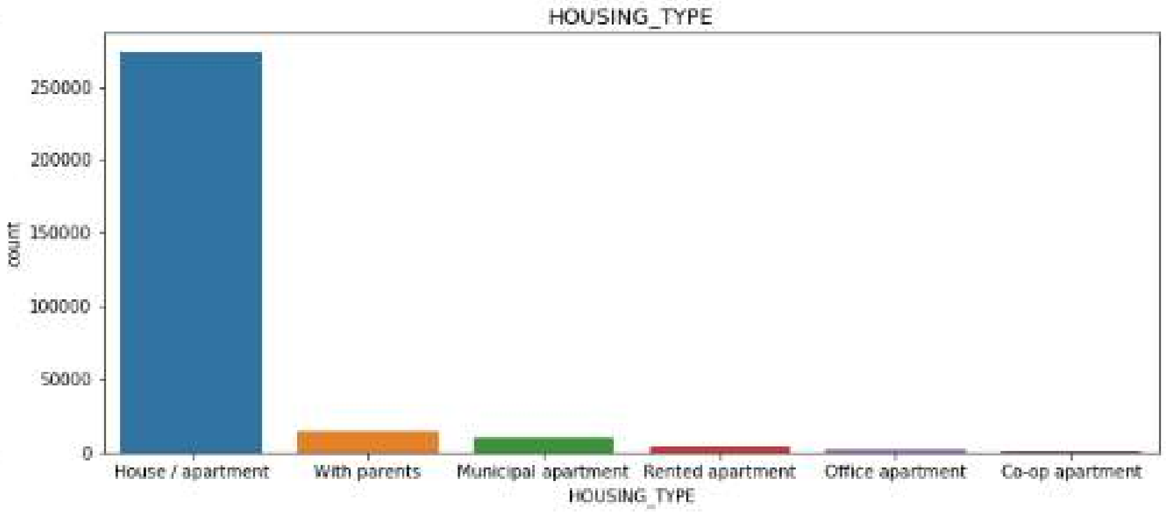


Categories with lowest quantities can be combined with other category.

VISUALIZATION

Data Exploration: Income type and housing type.

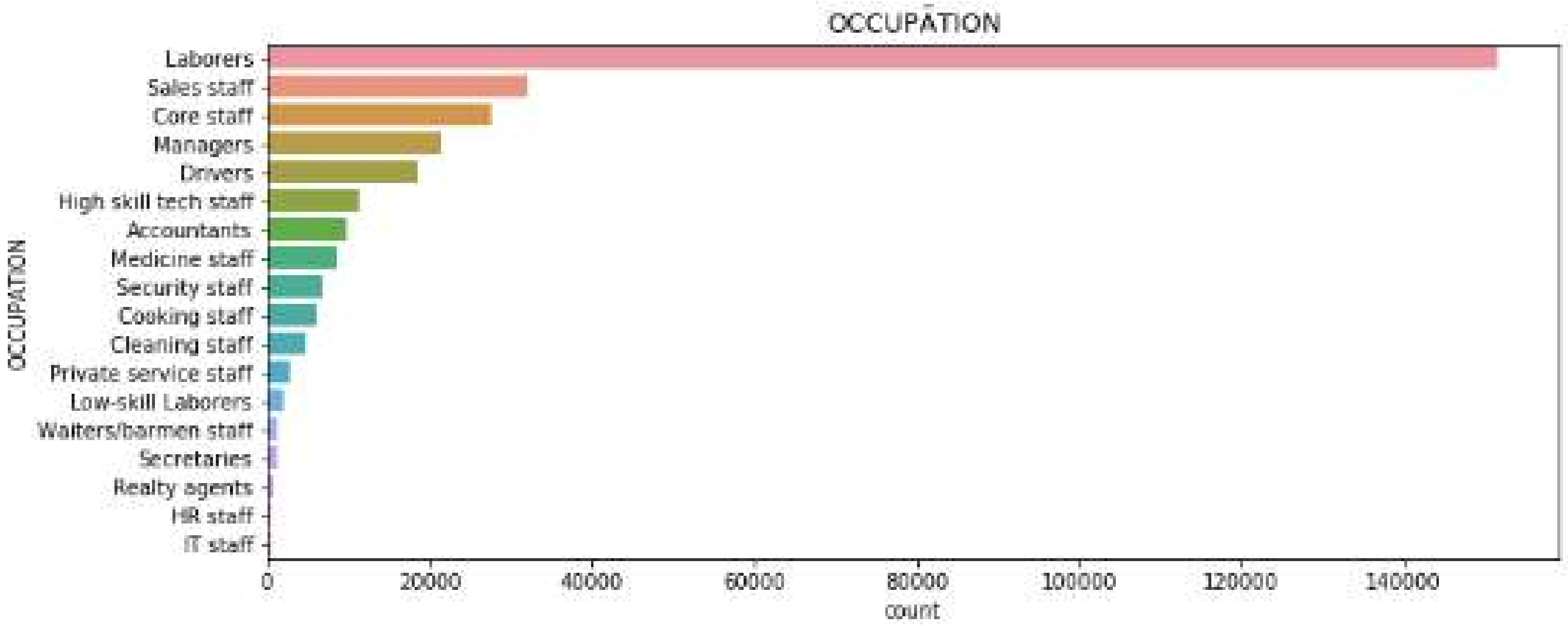




Categories with lowest quantities can be combined with other category.

VISUALIZATION

Data Exploration: Occupation type.

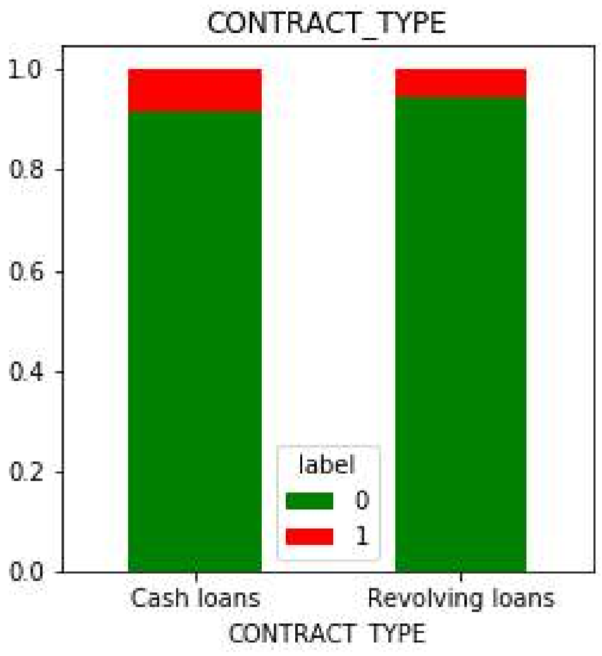


Almost of clients work as Laborers.

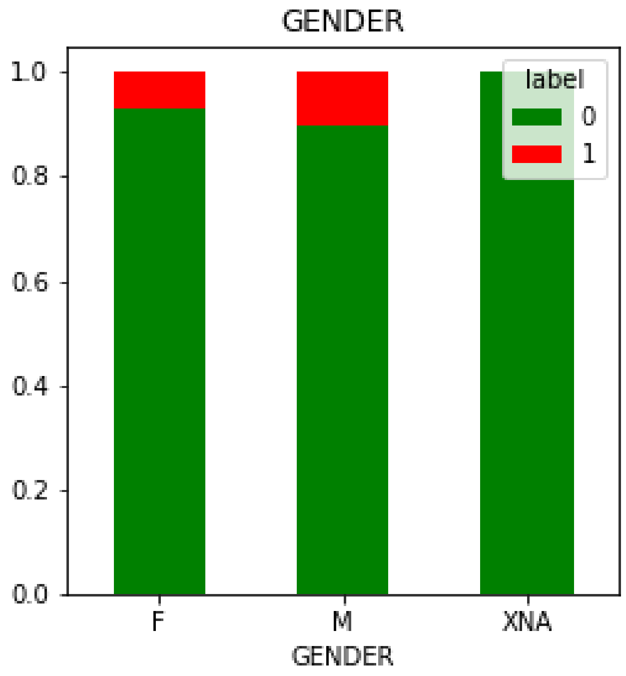
HOME CREDIT

VISUALIZATION

Data Exploration: Contract type and Gender VS label



Proportion label 0 and 1 in categories N and Y same, around 0.9 and 0.1.

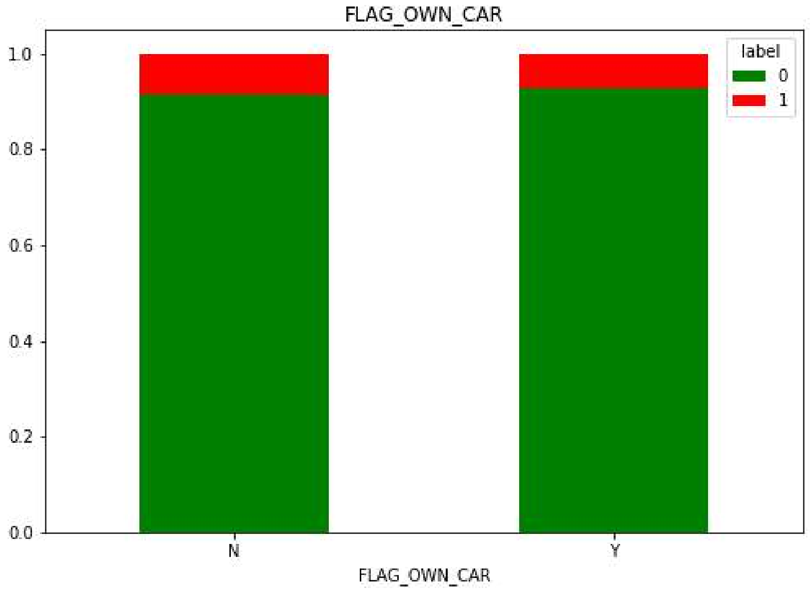


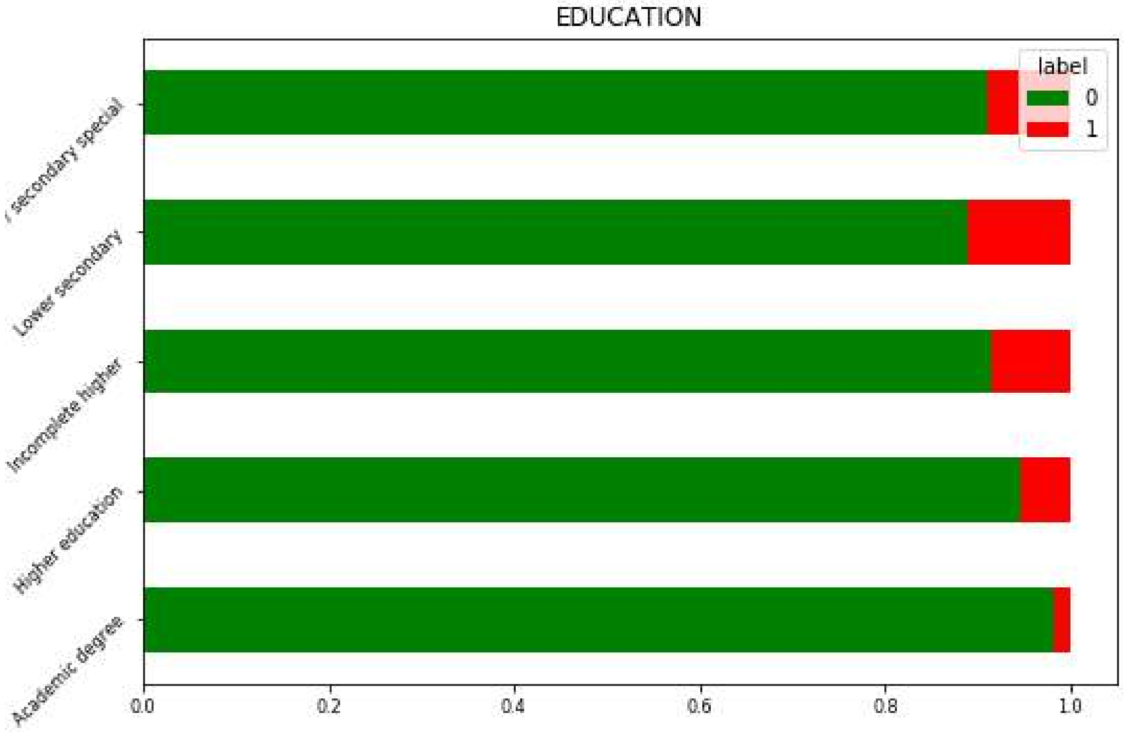
Category XNA in GENDER labelled by 0.

Plot shows proportion per label. Proportion around 0.9 for label 0 and 0.1 for label 1.

VISUALIZATION

Data Exploration: Flag_own_car and education VS label



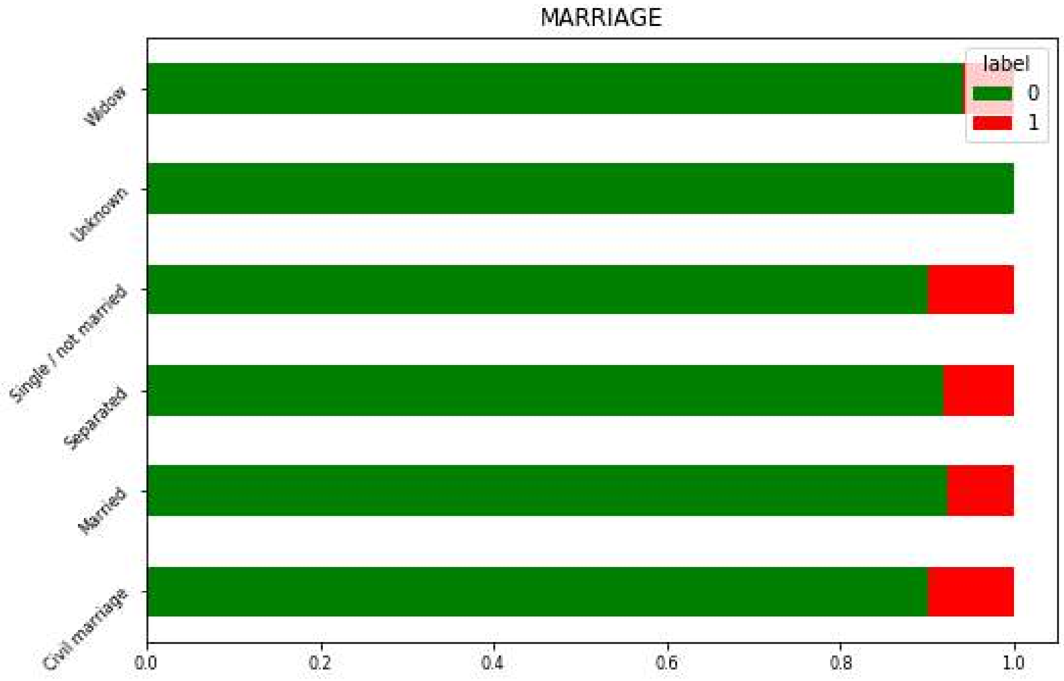


Proportion label 0 and 1 in categories N and Y same, around 0.9 and 0.1.

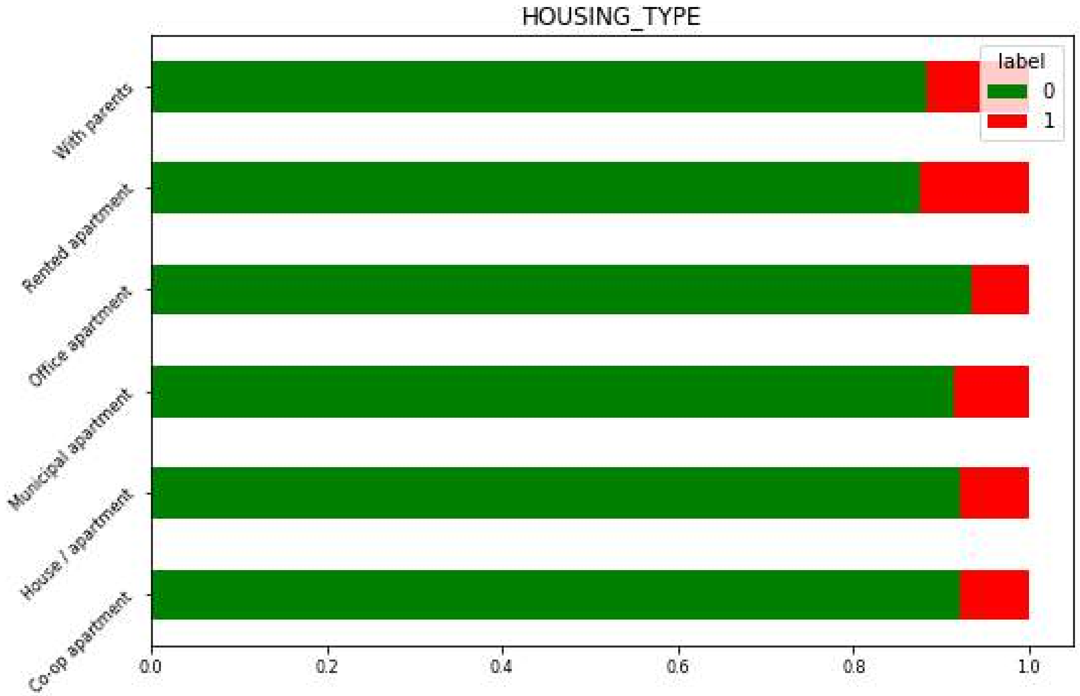
Label 0 in Academic degree higher than label 1.

VISUALIZATION

Data Exploration: Marriage and housing_type VS label



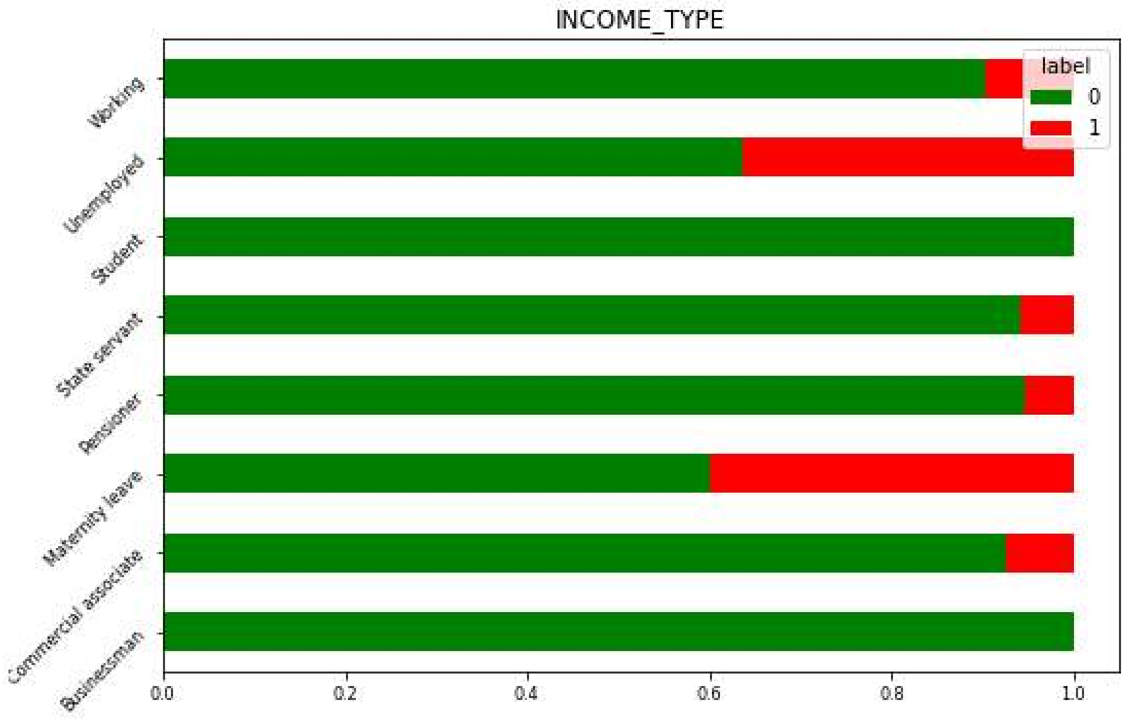
Unknown category in Marriage labelled by 0.



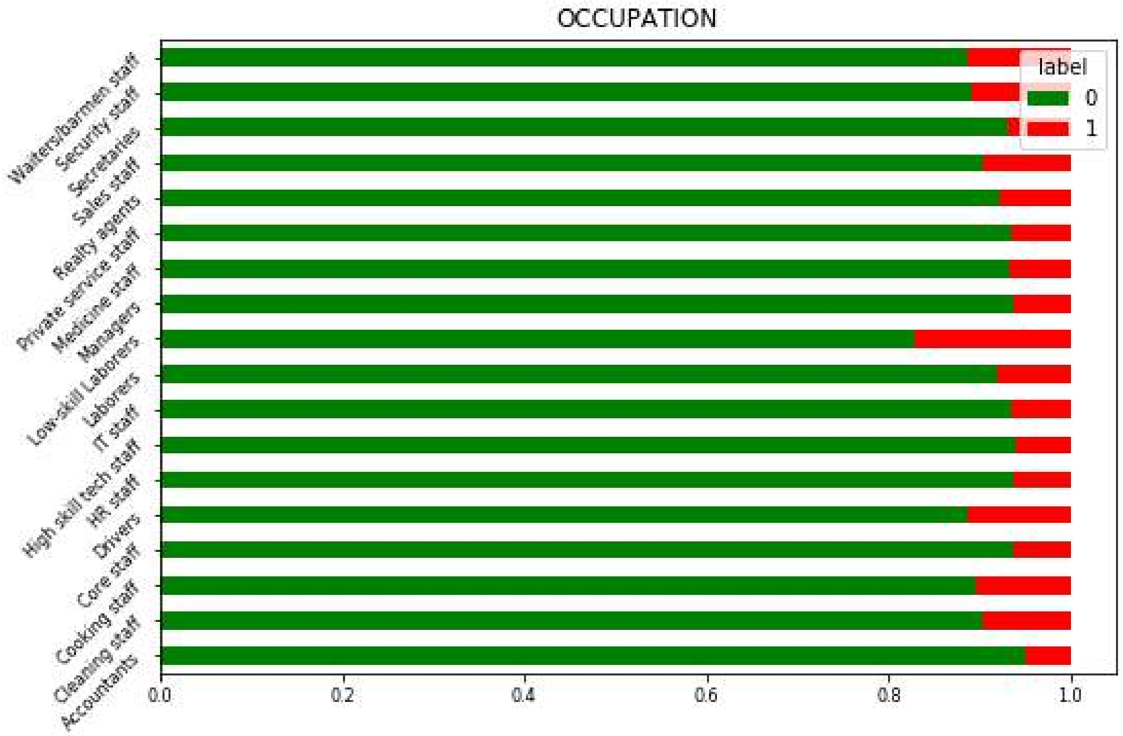
Trend proportion label 0 and 1 in each categories around 0.9 and 0.1

VISUALIZATION

Data Exploration: Occupation and income type VS label



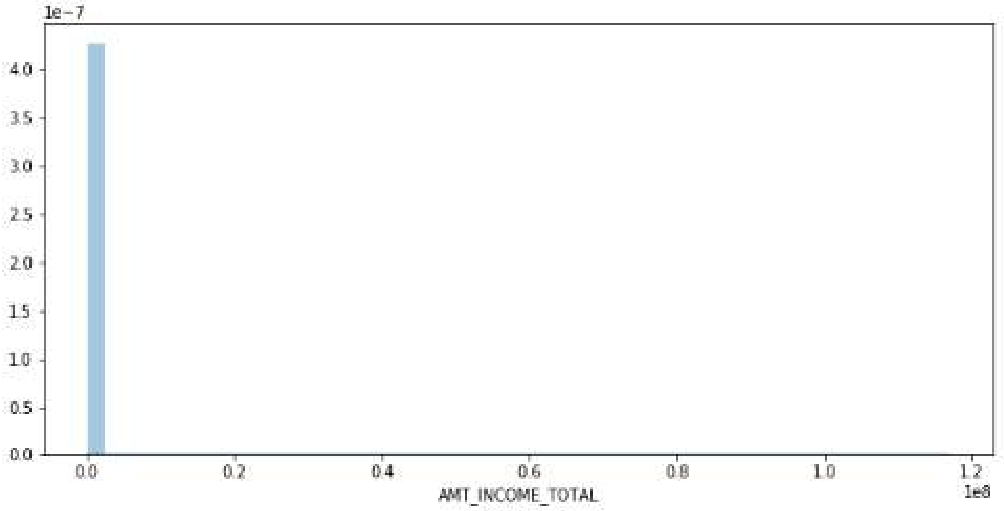
Businessman category labelled by 0. Maternity leave and unemployed have highest label 1, around 0.6

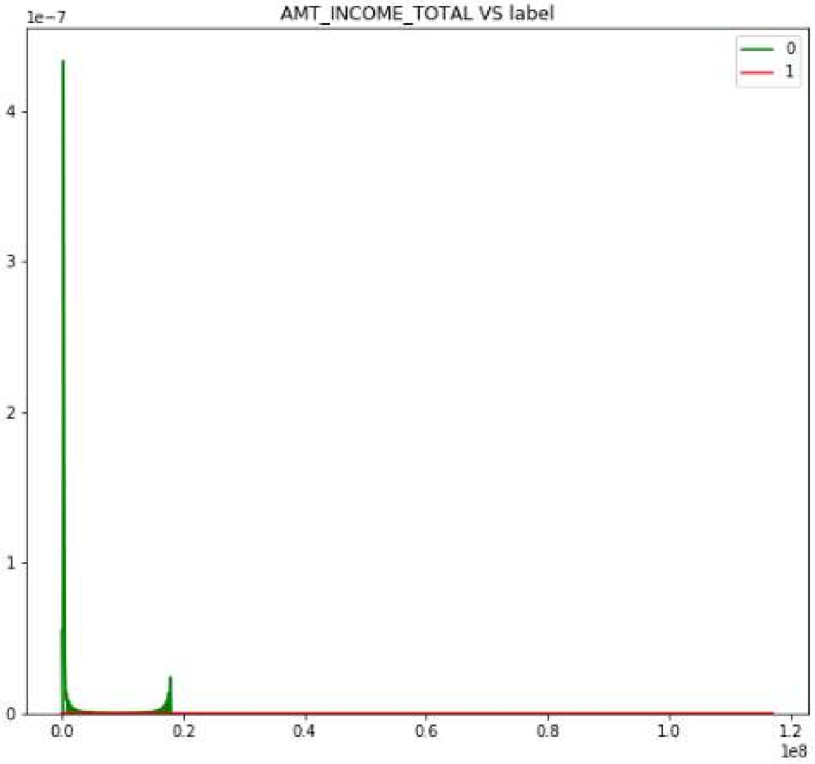


Only low-skill laborers category has highest proportion on label 1.

VISUALIZATION

Data Exploration: Amount total income VS label distribution.

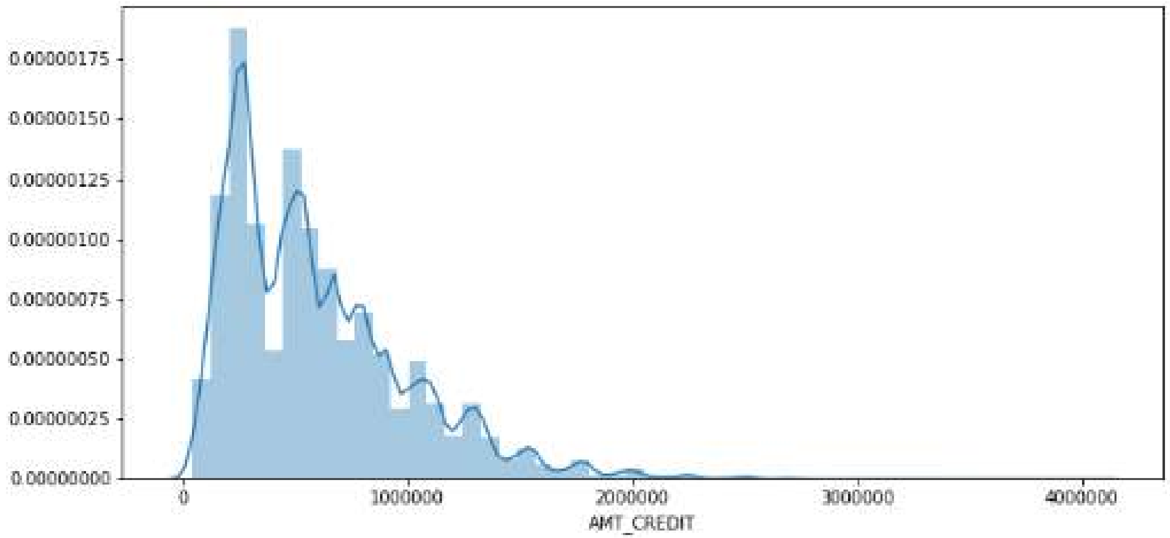


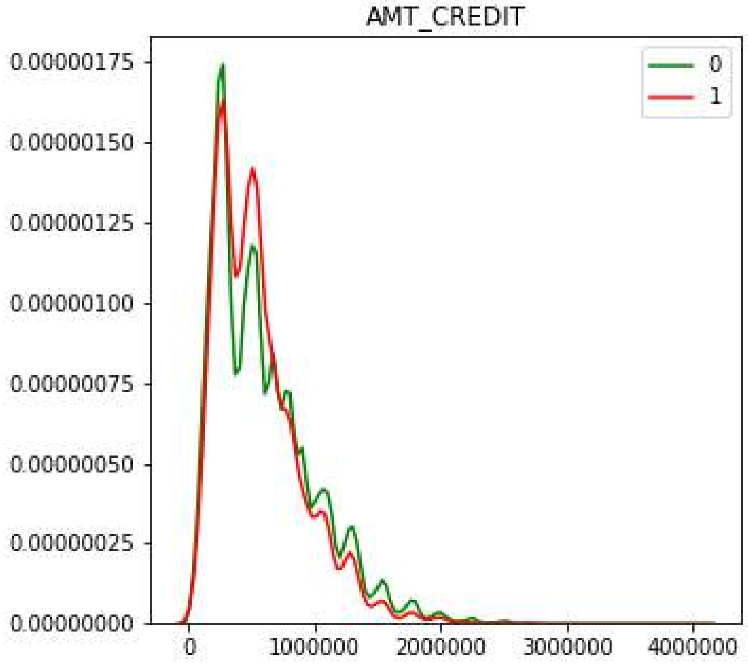


Maximal value from AMT_INCOME_TOTAL very large.

VISUALIZATION

Data Exploration: Amount credit VS label distribution.



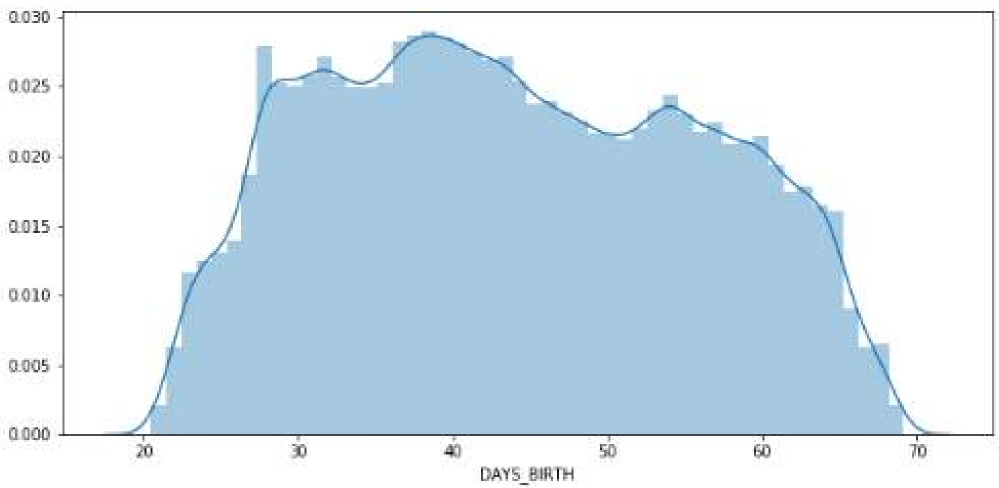


Client's AMT_CREDIT has maximal value very large. And decrease at around 1000000

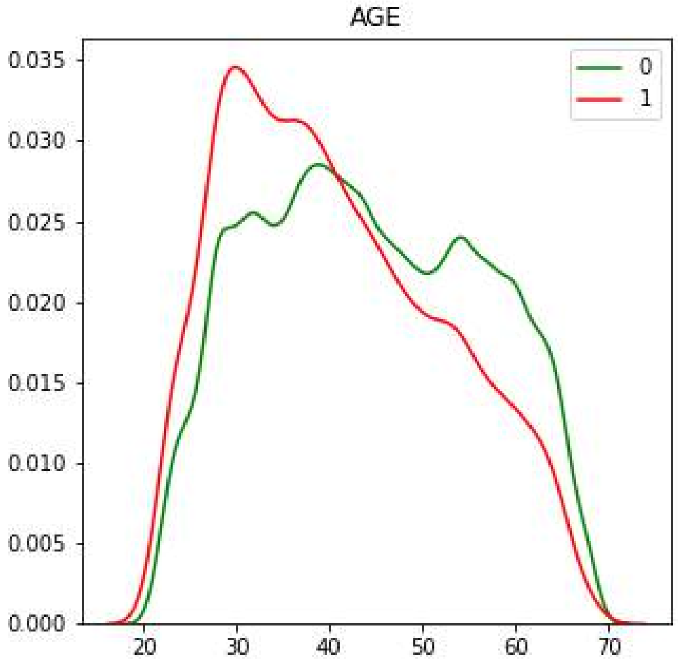
Clients who have difficulty to repay loan, have amount of credit lower than client who will repay loan on time.

VISUALIZATION

Data Exploration: Age VS label distribution.



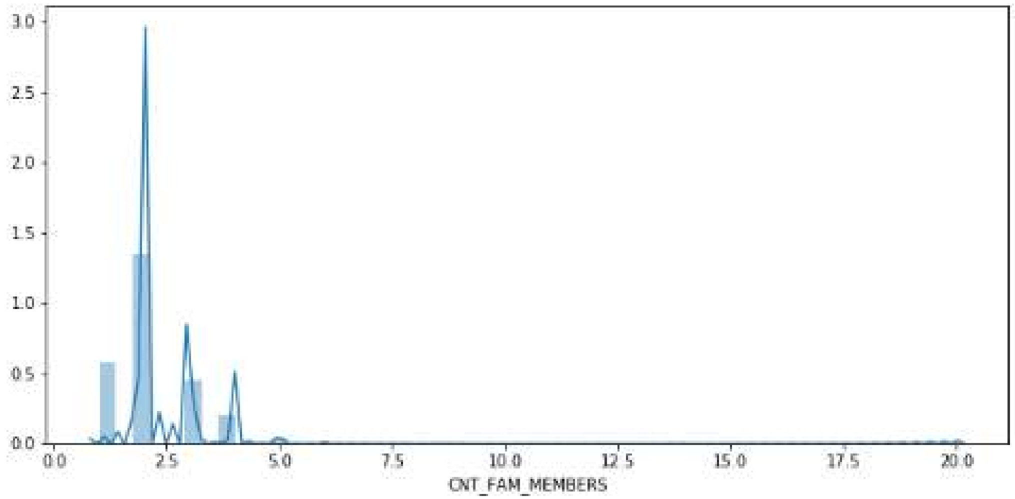
Range of client's age is around 20 until 68 years old.



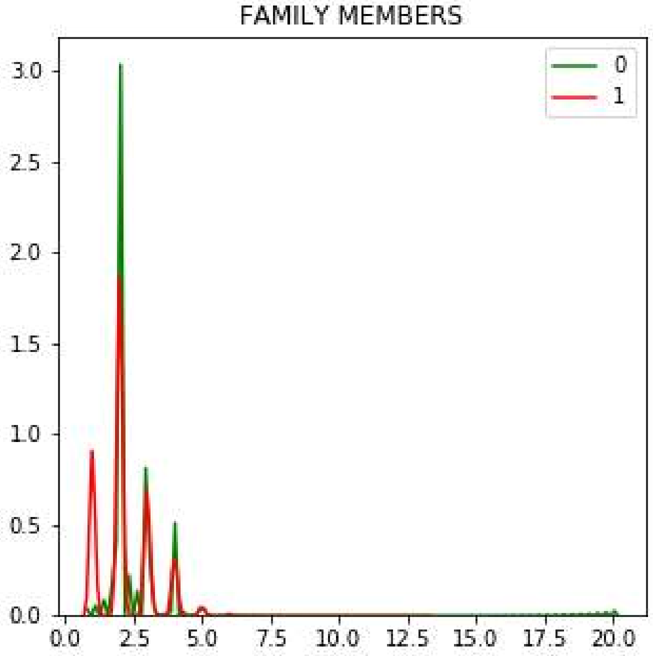
Clients who have difficulty to repay loan have range of age between 20 until 41 years old.

VISUALIZATION

Data Exploration: Number of family members VS label distribution.



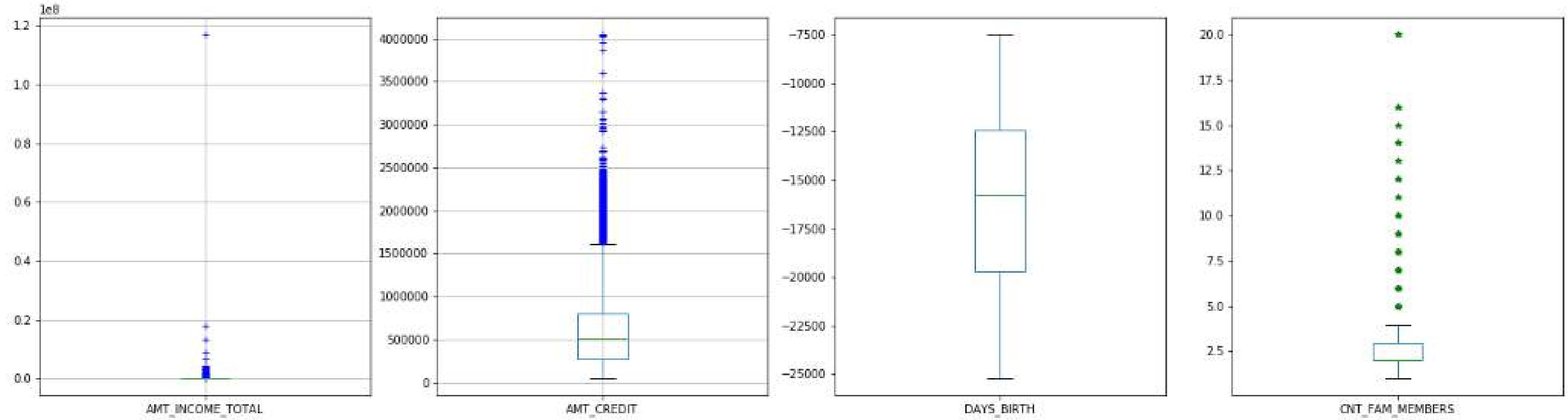
Number of family members dominant at two.



Client who have difficulty to repay loan have family members lower than clients who will repay loan on time.

OUTLIER

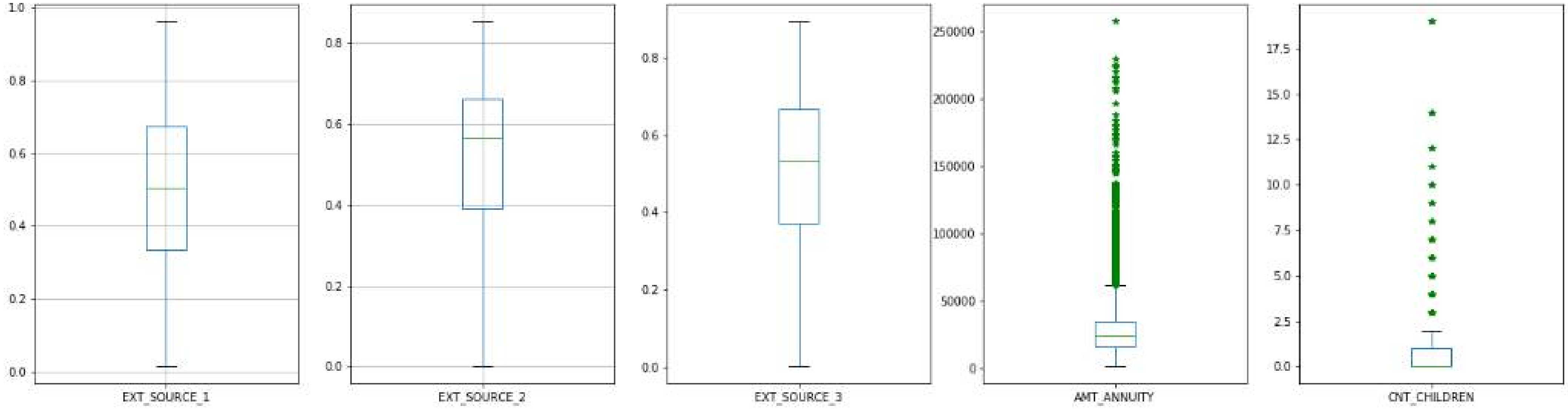
Data Exploration: Check outlier for numerical variables



Only DAYS_BRITH that has no outlier.

OUTLIER

Data Exploration: Check outlier for numerical variables



AMT ANNUITY and CNT_CHILDREN have an outlier (anomaly data).

OUTLIER HANDLING

Methods to handle outlier:

- Remove the observations,
- Replace with value of upper side or lower side

Calculate IQR, Q1, Q2, Q3 and upper side (Q3 + (1.5*1QR))

Outlier: Replace with value of upper side

AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, CNT_CHILDREN, CNT_FAM_MEMBERS

MODELLING

Three algorithm that used are:

- Logistic Regression

  Logistic regression used logit function in prediction the probability.
- Decision Tree

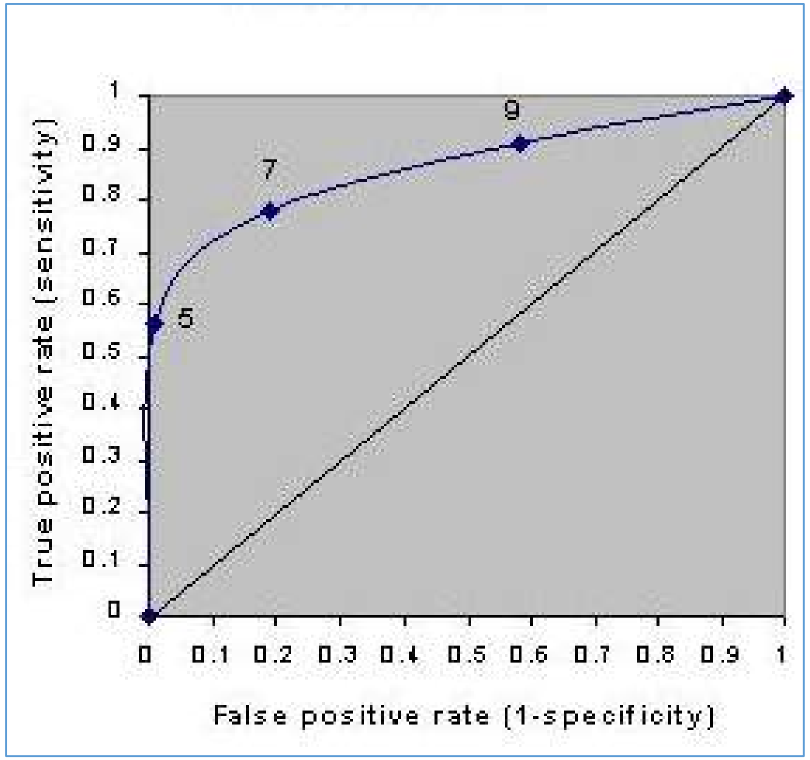
  This algorithm will find the most significant independent variable to create a group.
- Random Forest

  This algorithm build multiple decision trees and merges them together and use bagging method.

MODEL EVALUATION

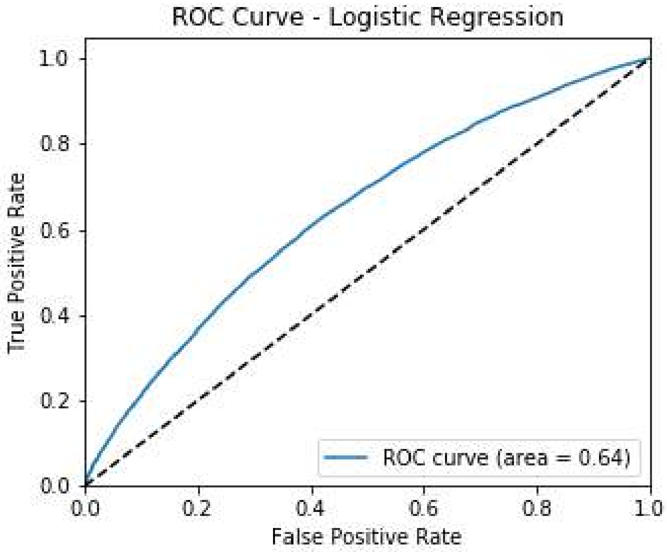
ROC (Receiver Operating Characteristic)

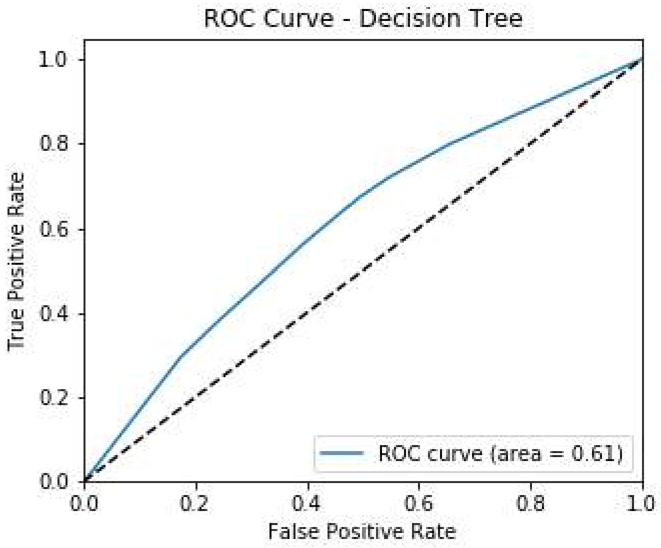
- The graph shows the true positive rate versus the false positive rate.
- This metric is between 0 and 1 with a better model scoring higher.

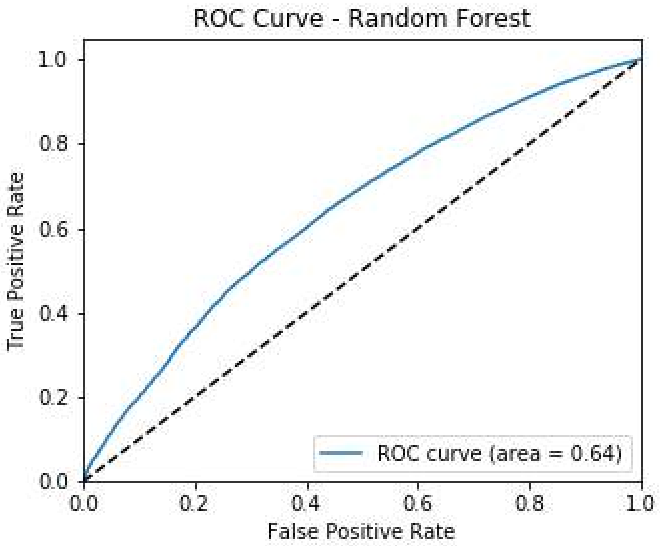


EXPERIMENT I & RESULT

EXPERIMENT I : Training dataset with selected columns







	prediction_label	0	1
0	0.0	84693	7317

Accuracy : 0.92

But BAD Model!!

FUTURE ENGINEERING

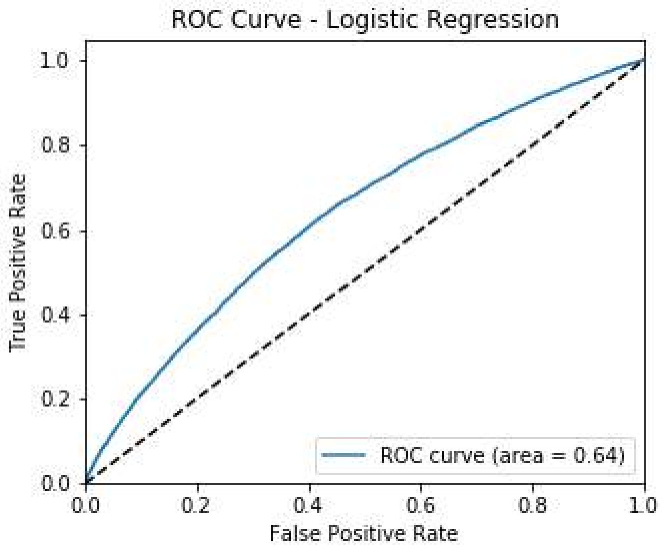
Let’s do some future engineering!

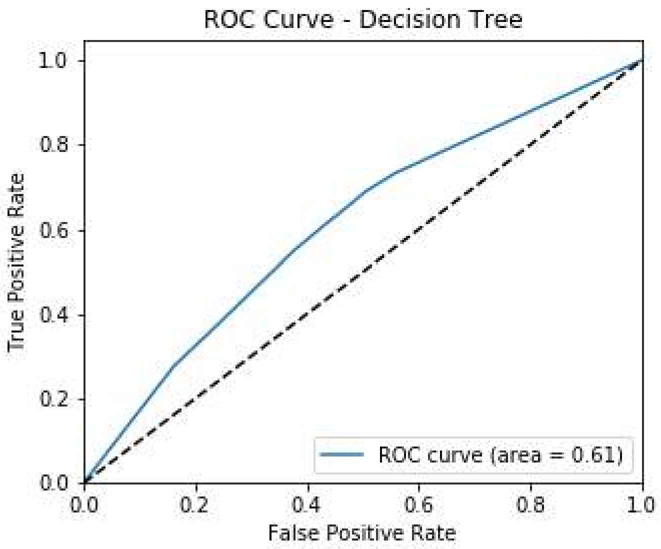
- ✓ Handle outlier,
- ✓ Handle the lowest categories,
  ex: Convert XNA category in GENDER variable to F.

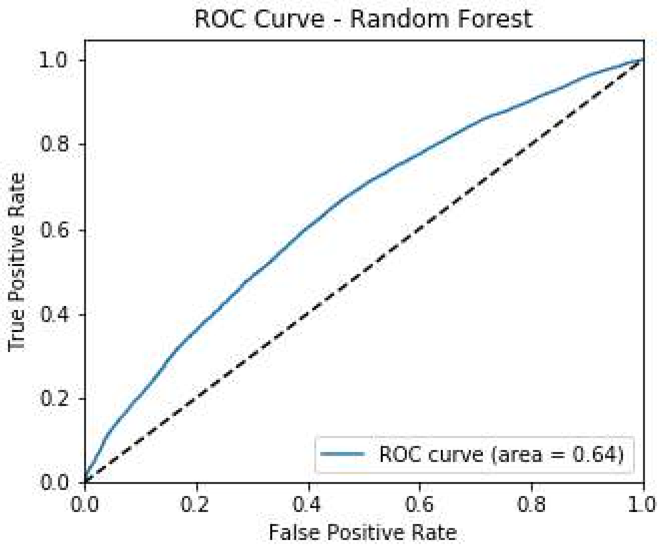
- ✓ Convert DAYS_BIRTH to the years, and
- ✓ Handling of imbalance data : Down sampling (70:30)

EXPERIMENT II & RESULT

EXPERIMENT II: Do Down sampling and some future engineering







No significant increase, we still have ROC curve around 0.6 for those three models.

FUTURE ENGINEERING II

Let's do some future engineering!

- ✓ Handle outlier,
- ✓ Handle the lowest categories,
  ex: Convert XNA category in GENDER variable to F.

- ✓ Add some variables : CNT_CHILDREN,AMT_ANNUITY, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3
- ✓ Handling of imbalance data : Down sampling (70:30)

Heatmap

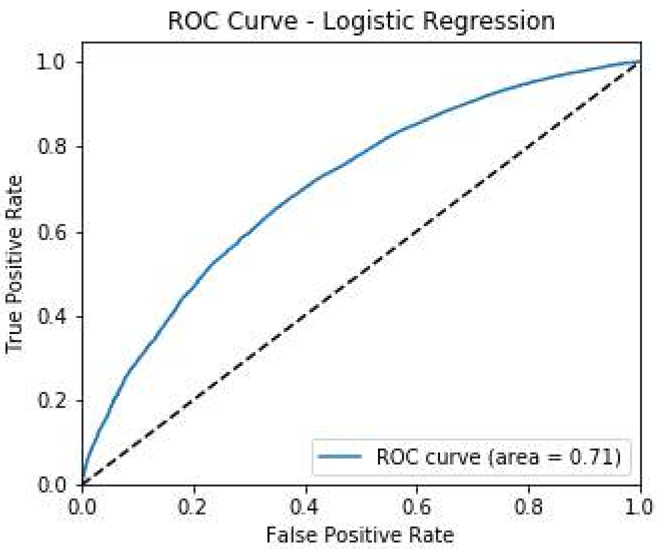
Strong correlation with label:

- ✓ EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3
- ✓ DAYS_BIRTH

HOME CREDIT

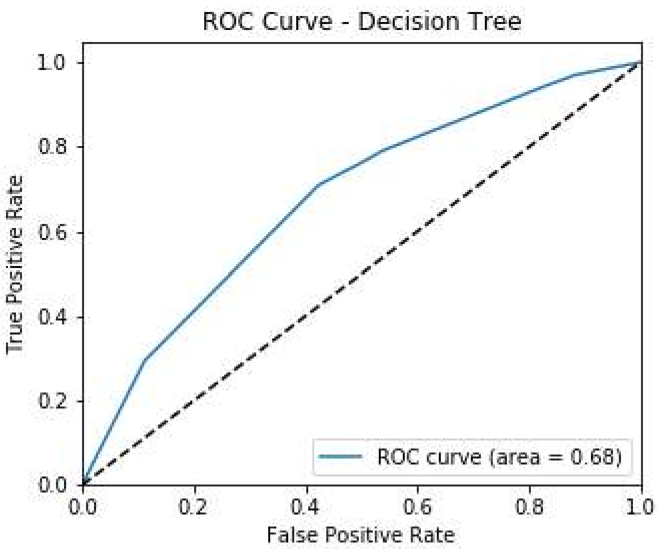
EXPERIMENT III & RESULT

EXPERIMENT III: Add some variable (CNT_CHILDREN,AMT_ANNUITY, EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3) and do a down sampling (70:30).



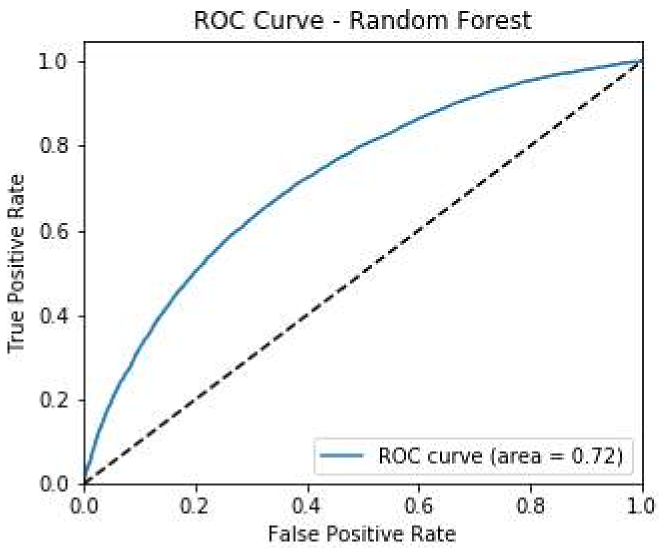
prediction_label		0	1
0	1.0	177	443
1	0.0	16879	7040

```
Accuracy = 0.71
Sensitivity = 0.06
Specificity = 0.99
Precision = 0.71
```



prediction_label		0	1
0	1.0	911	1244
1	0.0	16145	6239

```
Accuracy = 0.71
Sensitivity = 0.17
Specificity = 0.95
Precision = 0.58
```



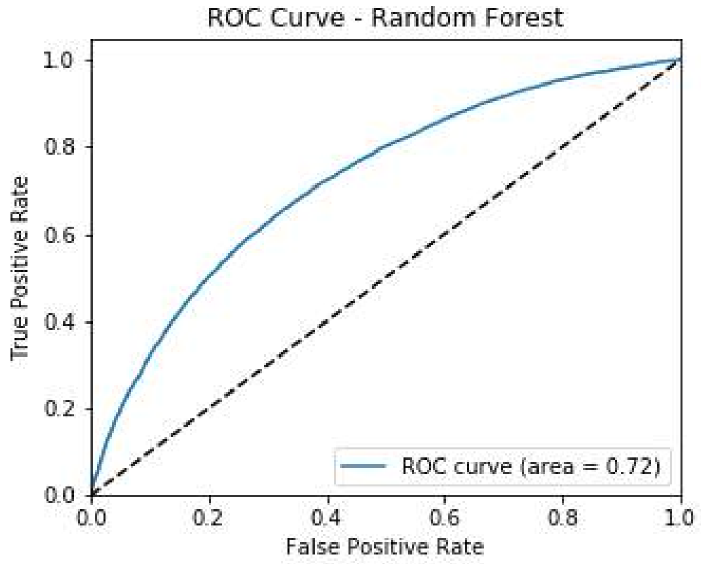
prediction_label		0	1
0	1.0	231	535
1	0.0	16825	6948

```
Accuracy = 0.71
Sensitivity = 0.07
Specificity = 0.99
Precision = 0.70
```

EXPERIMENT III & RESULT

Better Model!!

This model (rfModel_d.transform) will be used to make a prediction.



	prediction_label	0	1
0	1.0	231	535
1	0.0	16825	6948

```
Accuracy = 0.71
Sensitivity = 0.07
Specificity = 0.99
Precision = 0.70
```

IMPLEMENTATION

In implementation model to data test, it should be noticed that **every single steps done** on data **train** also should be **done** on the data **test**.

In this case:

- ✓ **Columns selection**
- ✓ **Rename columns**
- ✓ **Fill missing value**
- ✓ **Handle outlier**
- ✓ **Handle lowest category**

Model used : **Random Forest (rfModel_d.transform)**

application_test.csv

HOME CREDIT

CONCLUSION

- ✓ Imbalance classes could be dangerous because it’s predicting only 1 class, in this case only predicting 0.
- ✓ Accuracy is not best metric. Despite we have 0.92 accuracy.
- ✓ Use AUC ROC as a metric.
- ✓ Use down sampling to handle imbalance classes.
- ✓ AUC ROC increase around 0.07 after down sampling and futures selection.
- ✓ Random Forest more accurate doing the test than other models.
- ✓ With precision 70%, means around 535 client predicted difficult to repay.

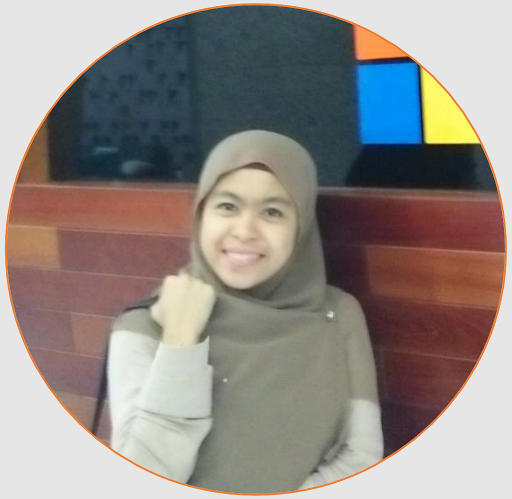
Step Further:

- ✓ Try over sampling,
- ✓ Try advance ML algorithm,
- ✓ Perform futures engineering,

Kaggle Score

0.715

Thank you



Syifa Silfiyana S

syifa.silfiyana@gmail.com

https://www.linkedin.com/in/syifa-silfiyana