Predicting Credit Default - Home Credit

(Data Science Intensive Capstone Project)

By Rohan Agale

Mentor - Ben Bell

Reviewer - Blaine Bateman

Executive Summary

Problem Statement

Home Credit is in the business of providing consumer credit to financially underserved population. At the time of credit application, it wants to **predict which clients will repay the loan and which ones will default on the loan.**

Data

The data contains information about **application, applicant's details, her credit history with Home Credit**. We also have **her credit history from the Credit Bureau**. Since data is from Kaggle, the data understanding is limited to data dictionary & discussions in the forum.

Insights

Attributes of customers with better credit repay capability :

1. Higher value of external factor 2 (Most important feature)
2. Females
3. Higher Education
4. Pensioners, State Servants.
5. On time payment of installment of previous loans

Modelling

1. **0.7826 ROC AUC** score was achieved by **XgBoost** model. Light GBM performs slightly worse than with 0.7814. As XgBoost is 1.8 times faster than Light GBM in prections, we choose XgBoost as the best model.
2. XgBoost generalizes well on test data with ROC AUC of 0.7841.

Profit Maximization

Aiming for profit maximization, we choose the optimum threshold for giving the loans.

1. The book profit increases to **3.29%** from current profit of 1.49%
2. And the expected coverage of loan is 77% of the population.

Topics -

0. Objective
1. Data Wrangling
2. Exploratory Data Analysis
3. Feature Engineering
4. Modelling
5. Profit Maximization

0 - Objective

Objective

Context

Home Credit is an international consumer finance provider. It is on a mission to **provide a safe, simple & fast borrowing experience to unbanked and underserved population**. On this mission, Home Credit wants to explore and utilize the predictive power of ML to extend credit access to deserving population and avoid defaults.

Problem

Identify the potential clients as those

1. who can **repay** the loans and
2. who can **default** on loans

at the time of application. Predicting the severity amount of default is out of scope.

Why

Identifying loan paying capacity at the time of application is a **win-win situation for Home Credit and applicants** because it helps

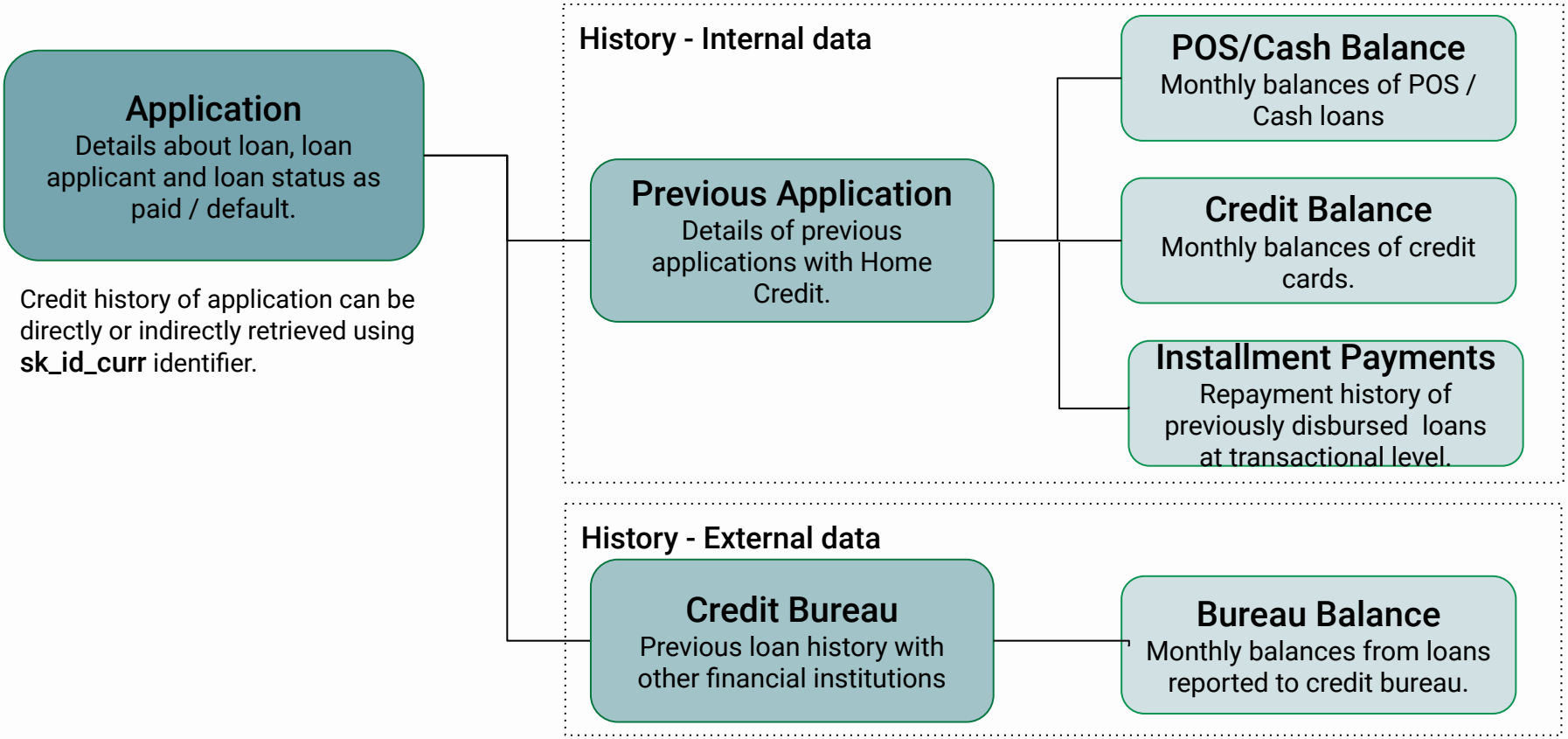
1. Home Credit to remain solvent by maintaining a healthy portfolio
2. Applicants to make an informed decision and avoid debt traps.

Stakeholders

1. Chief Risk Officer
2. Chief Finance Officer
3. Head of Analytics

1- Data Wrangling

Data Schematics



Data Specifications

#	Dataset Name	# Observations	# Variables	% of applications in the dataset
1	application	307,511	122	NA
2	prev_application	1,670,214	37	94.60%
3	pos_cash_balance	10,001,358	8	94.10%
4	credit_card_balance	3,840,312	23	28.20%
5	installments_payments	13,605,401	8	94.80%
6	bureau	1,716,428	17	85.70%
7	bureau_balance	27,299,925	3	35.70%
Total	-	58,441,149	218	-

94% of the applications had applied previously. And 94% r POS or cash credit in the past.

Only 28.2% of the applications had a history of credit card balances.

Monthly balances are not reported to Bureau for some credits.

As the data is from Kaggle, our understanding is limited to what is described in data dictionary and discussed in the Kaggle forum. The details can be found at the Data Description notebook & Data Wrangling notebook.

Data Wrangling - Combining the datasets

POS Cash balances, Credit balances and **Installment** are monthly or transactional details of each previous loan application. Hence, we summarize the experience of each loan, create new features and finally merge the dataset with application data.

1.Firstly summarize the experience of each loan using mean, maximum and most frequent functions so that we have one row per loan.

2. Create features from the last (most recent) loan experience as well because the applicant's recent financial well being may differ from her financial well being in the past.

Engineer new **behavioral and financial** features while summarizing the data. Eg

payment_delay - This feature indicates delay in payment from the due date of installment. This delay is counted in number of days. It is zero if the loan is paid before the due date.
in_sum_payment_delay - Sum payment_delay to get total delay for the loan.
rt_payment_delay - Normalize in_sum_payment_delay by total loan duration. This is the main feature we want.

1. Merge the summarized **POS Cash balances, Credit balances** and **Installment** datasets with **Previous Application** data. Let's call this prev_app_merged data.

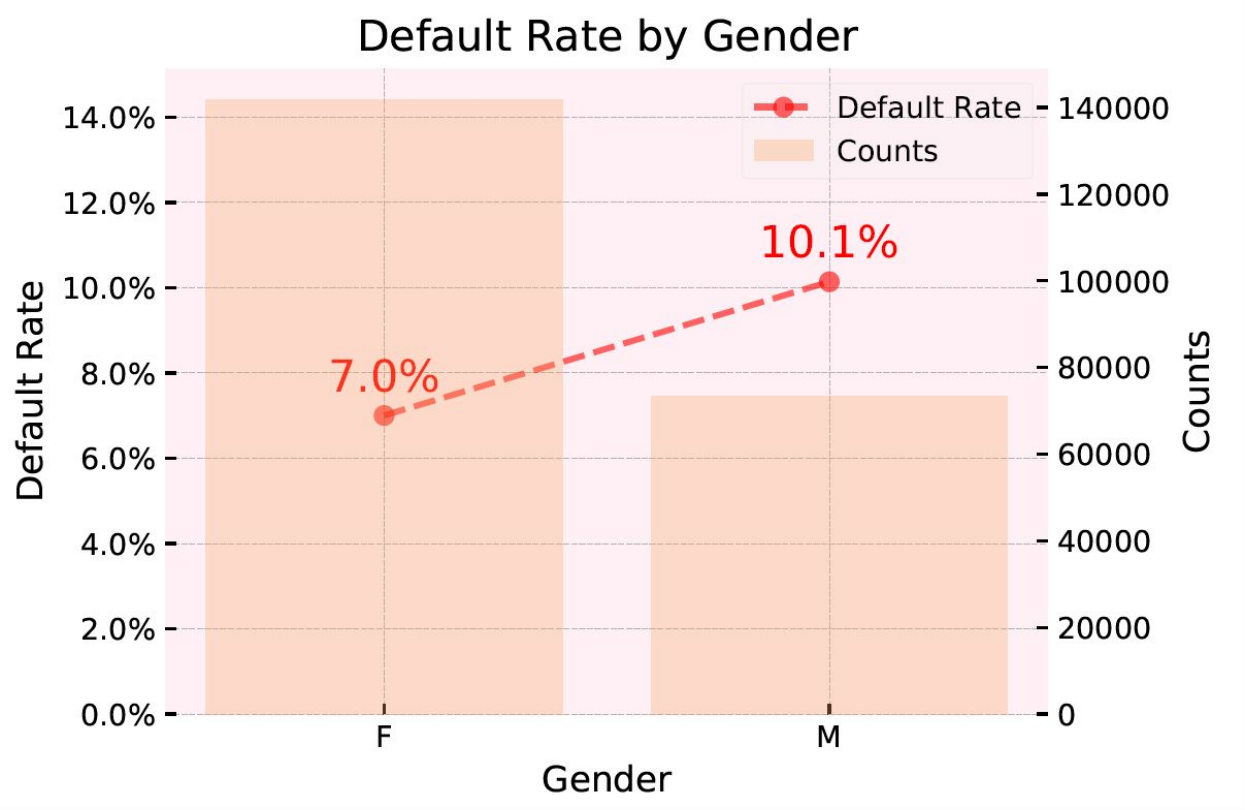
2. Secondly, summarize prev_app_merged so that we have one row per current application. We also create some new features. Finally, we merge the summarized prev_app_merged data with the **Application** dataset.

We follow same process for Bureau data. Summarize **Bureau Balance ->** Merge with **Bureau ->** Summarize **->** Merge with **Application)**

At the end, we have **307499 observations and 230 variables** including target.

2 - Exploratory Data Analysis

Gender : **Females have 3% points lower default rate.**



~65% of applicants are females.

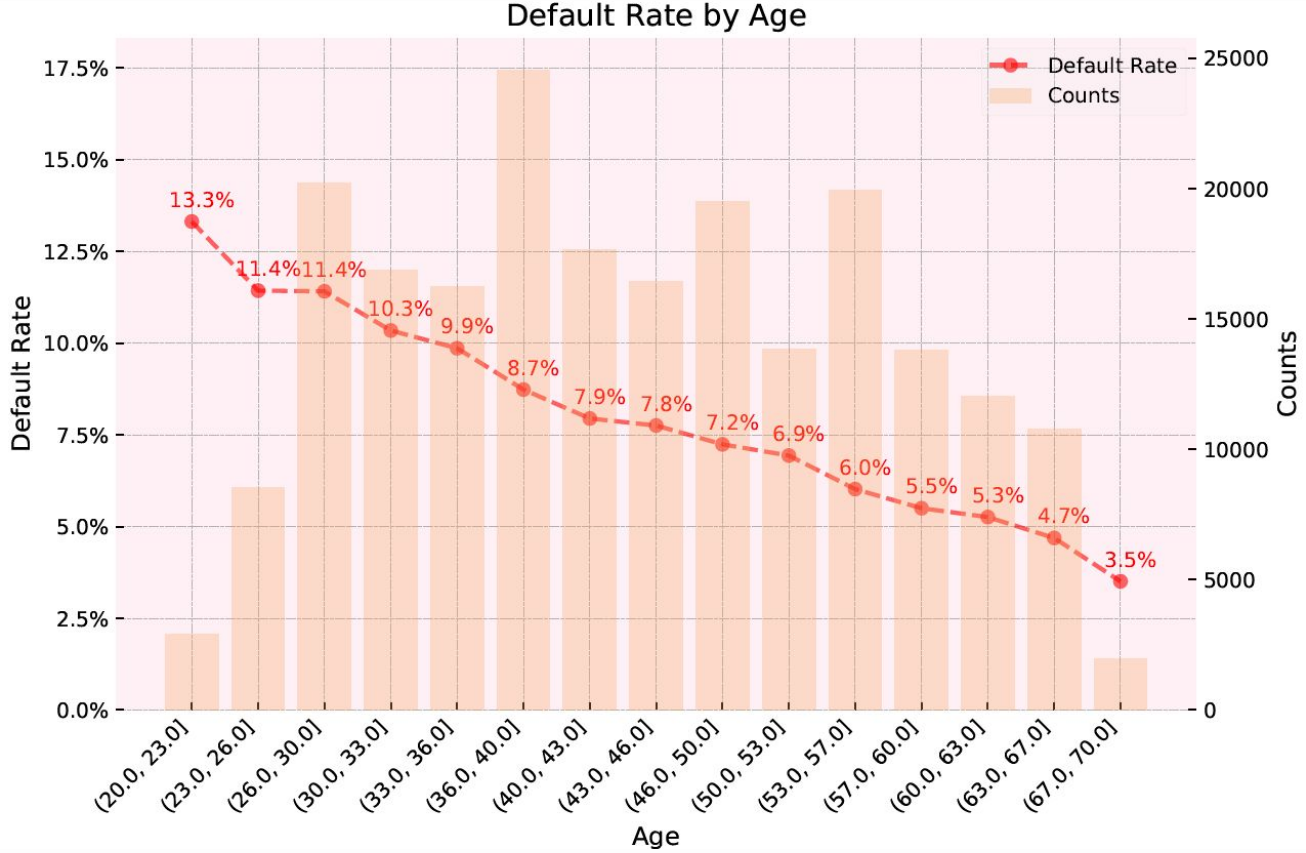
Females are on average 3% points better than males in paying back the loans. In general, females are conservating when it comes to managing finances and the trend is indicating the same.

** All these insights are derived from train data(70%).

Primary Y axis - Default rate(% default applications).

Secondary Y axis - Count of applications

Age : **Default rate decreases with age.**

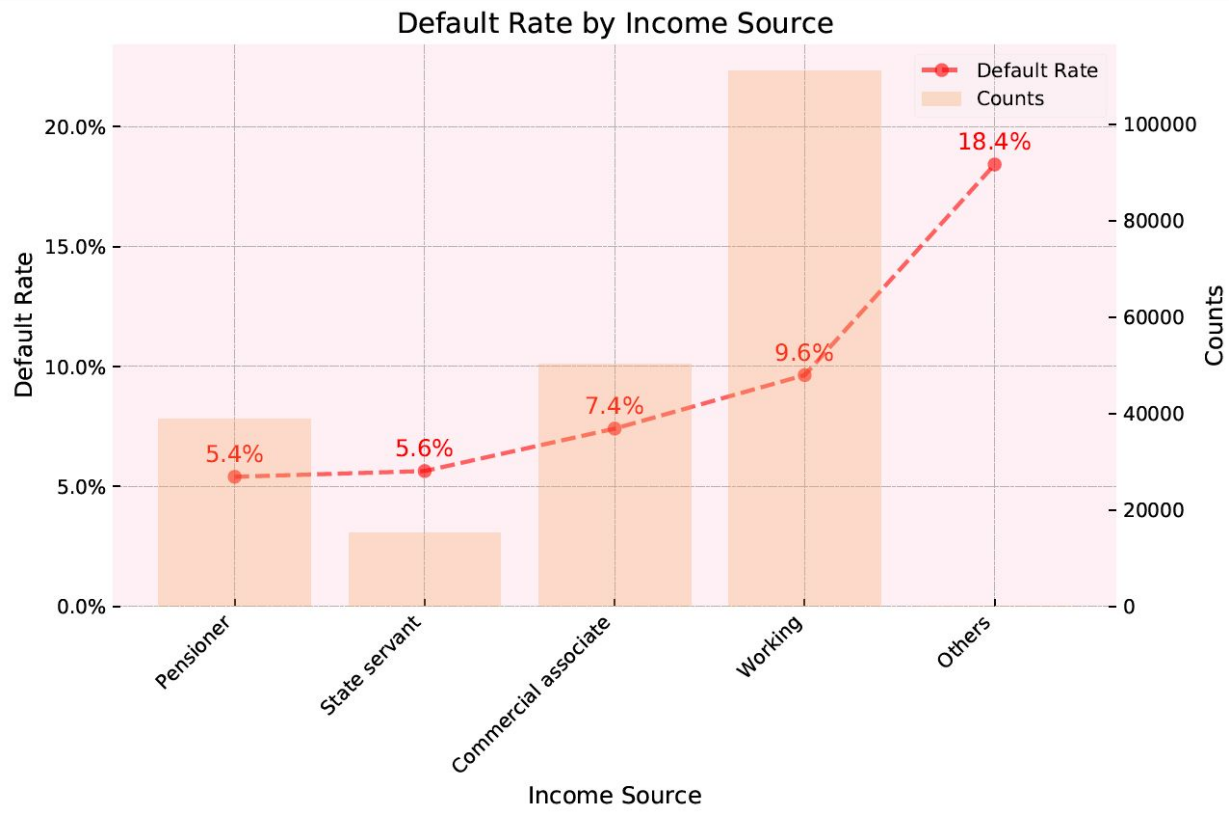


Applicant's age is within a reasonable range of 21-69. Home credit lends to clients across all ages.

We can clearly see that the default rate decreases as age increases. Average risk differentiation between the 2nd youngest and the 2nd last oldest groups is more than 7% points.

Raw feature is days_birth. We convert it to age and bin it for trend in default rate & count of observations.

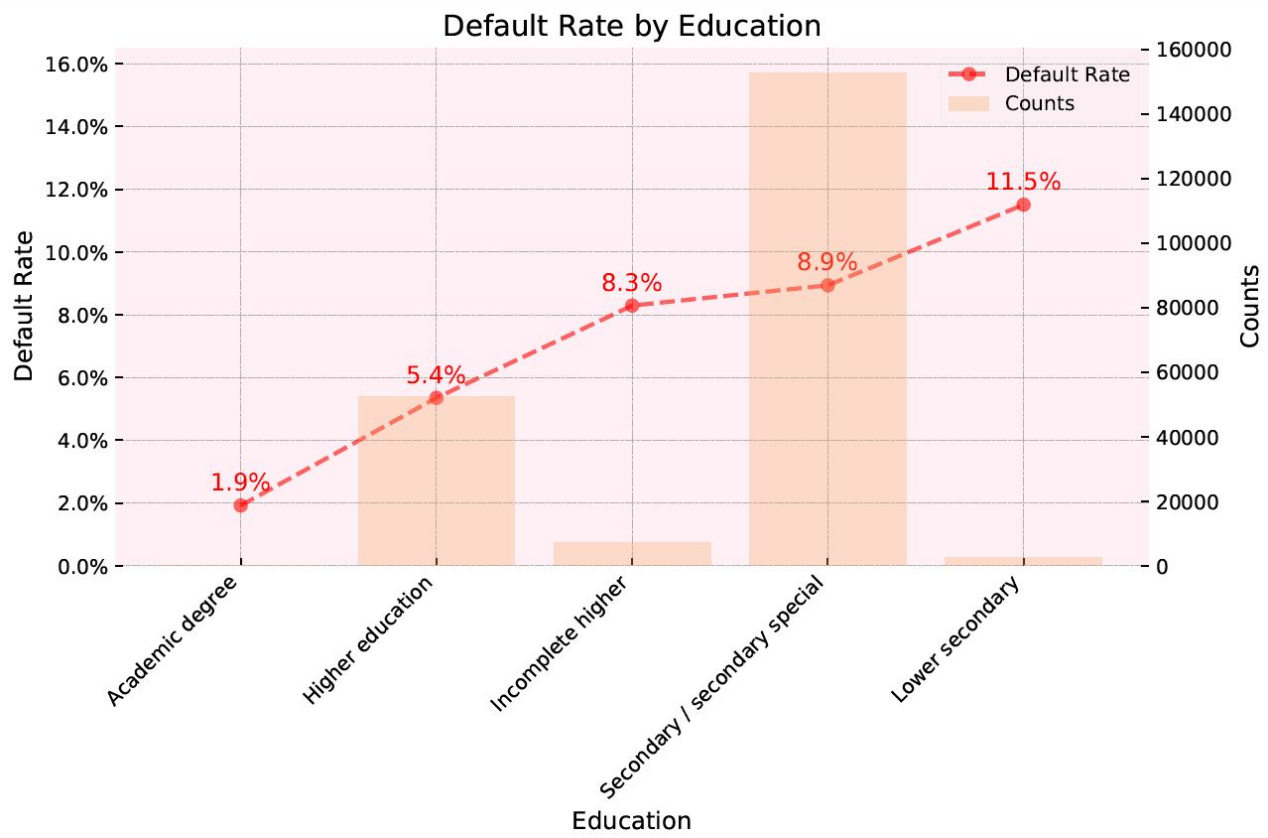
Income : **Pensioners & State servants are better at paying loans**



More than 50% loans are given to working professionals.

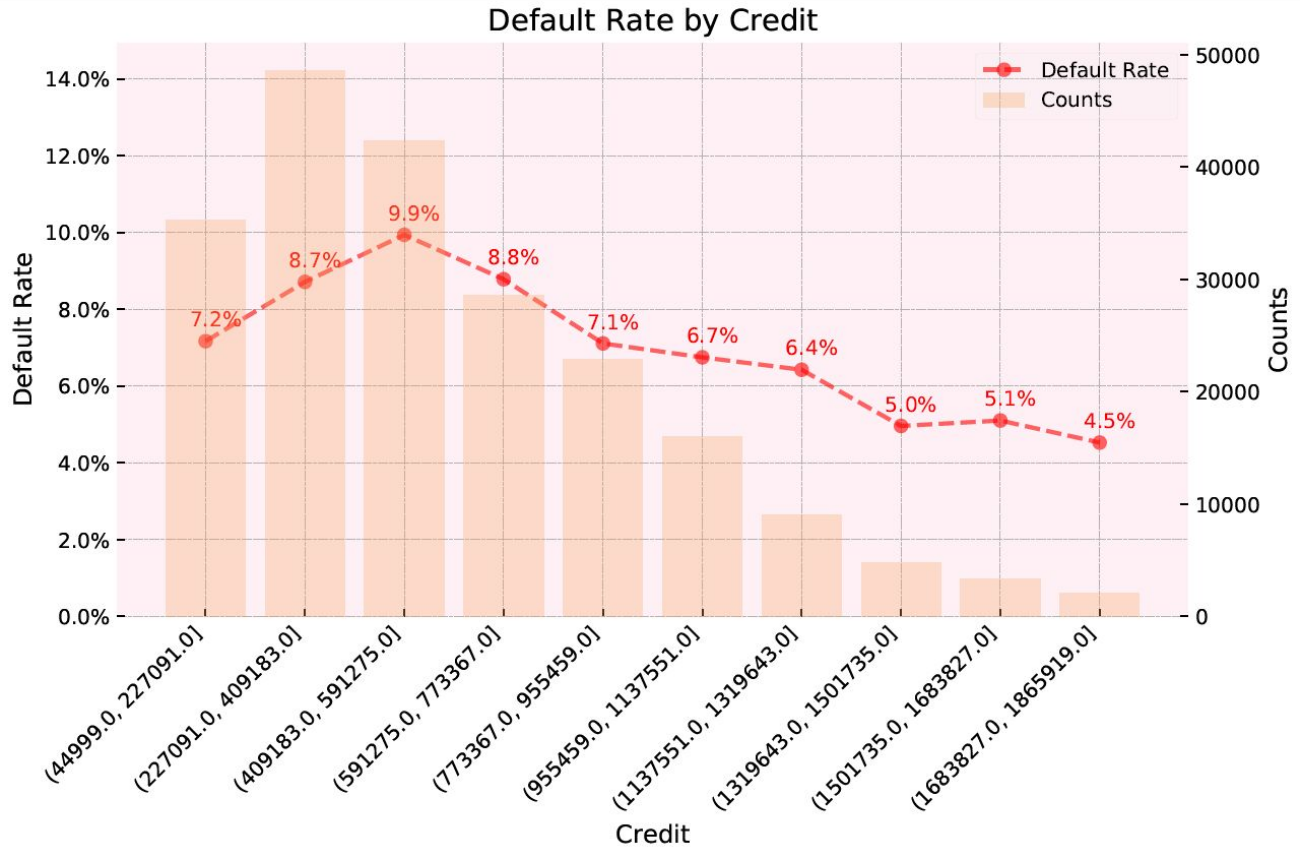
Pensioners and State servants are having much lower average default rate than working professionals.

Education : **Lower the education level, higher the default rate**



~70% of the clients fall into the secondary education category and ~25% of clients fall into the higher education category. Higher the education, lower the default rate.

Credit Amount - **Default rate peaks around median.**

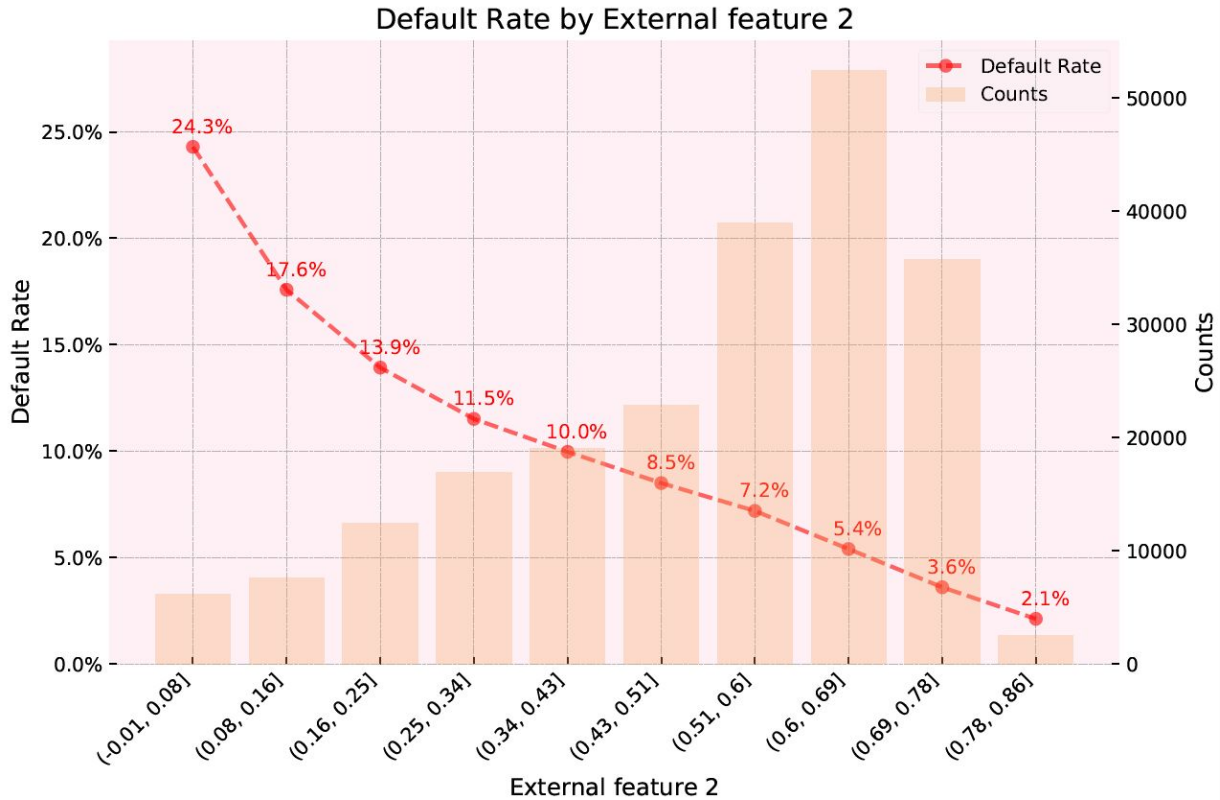


We can see an inverted V type trend. Initially, with an increase in credit amount, the default rate increases and it peaks around the median. Then it starts decreasing again.

For higher credit amounts, the underwriting process can be stricter than usual. We can ask business about the validity of these trends. We see **similar distribution and trends in goods price amounts.**

** Removed top 1 percentile for visualization purpose

External Feature 2 - **Decreasing trend with ~20% point risk differentiation**

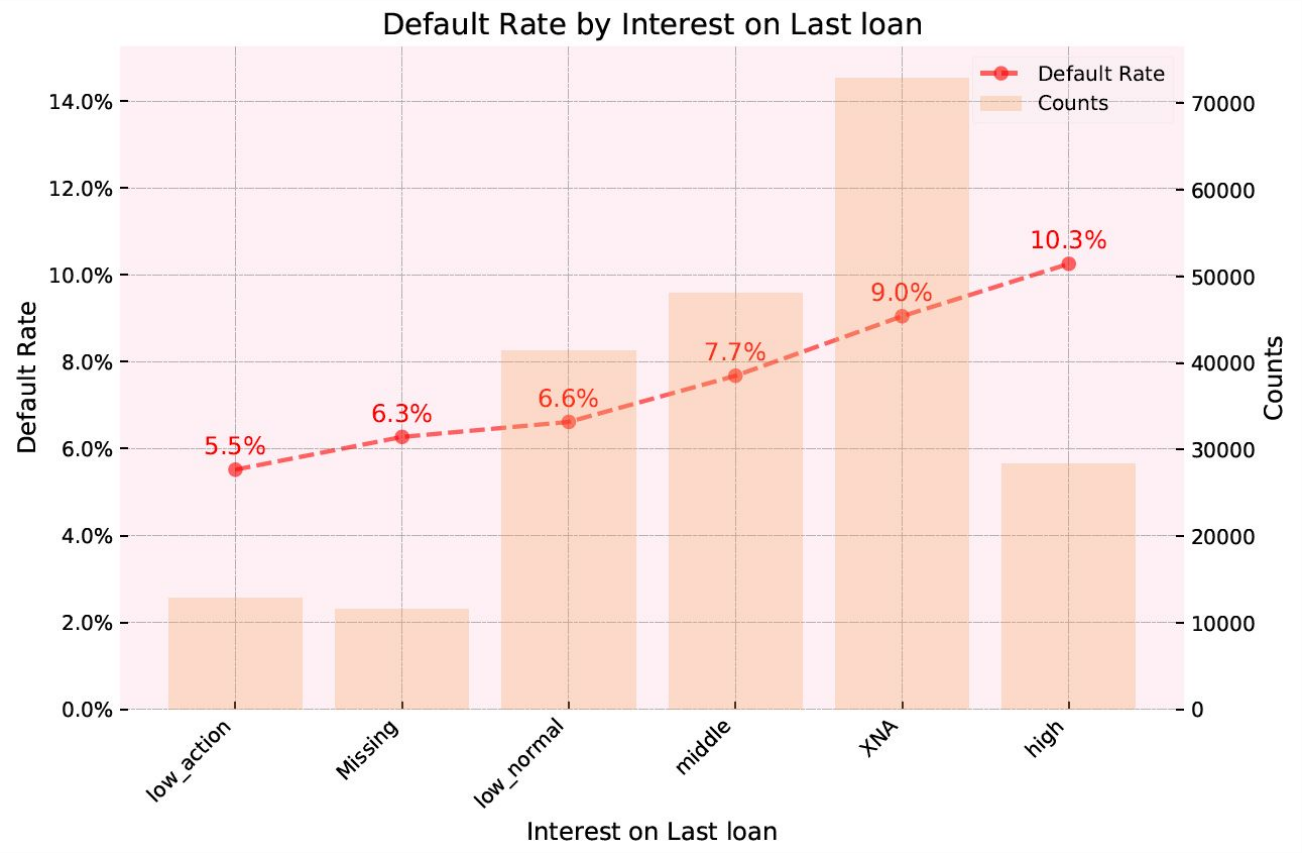


This normalized external feature is exhibiting a strong downward trend with excellent risk differentiation with >20% point difference between 1st & last decile.

We also have external features 1 & 3 with similar trends. But they have 56% and 20% missing values respectively. Whereas the external feature 2 has just 0.2% missing values.

** Removed top 1 percentile for visualization purpose.

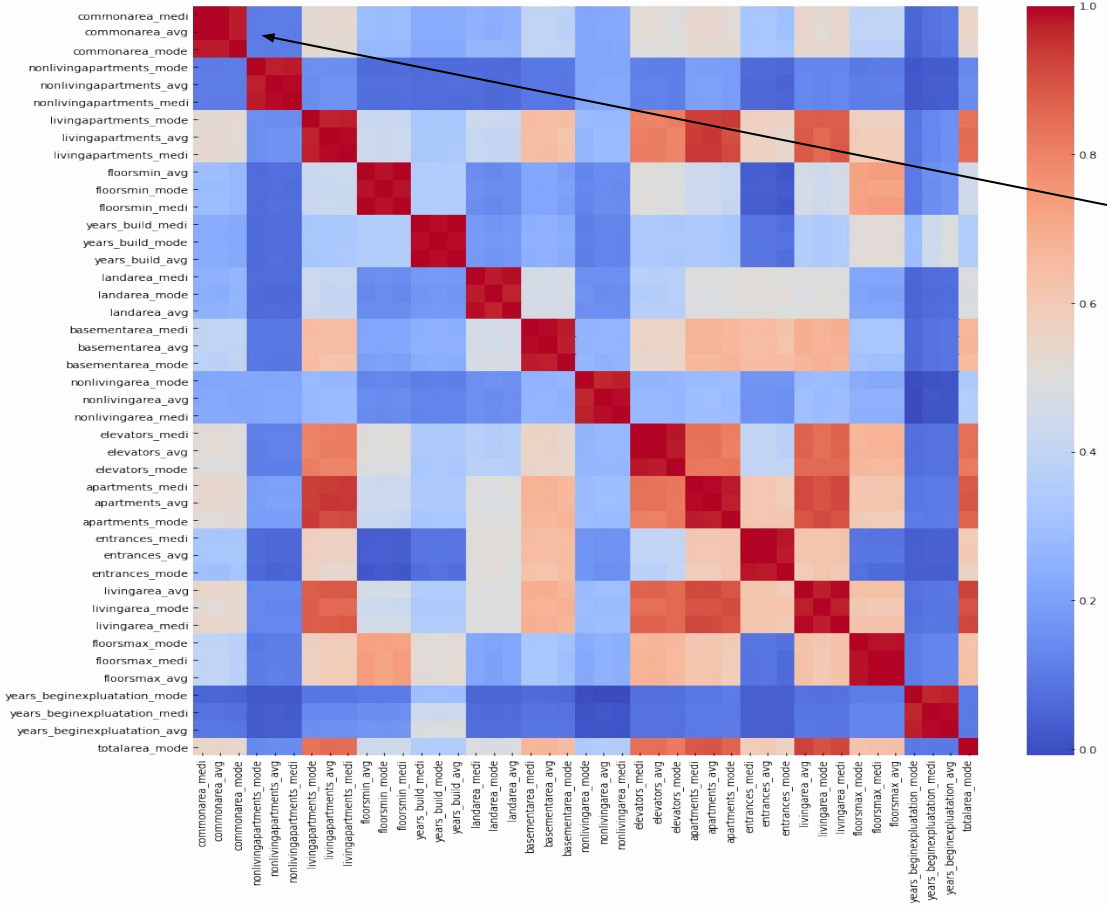
Interest on Last Loan - **Default rate is higher for higher interest rate**



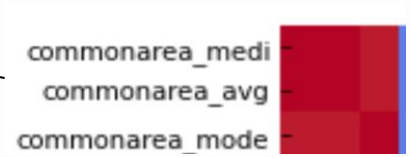
Default rate is higher for higher interest offered on the last loan. XNA interest rate is not defined in the data dictionary.

This is in some way a confirmation of the existing lending process.

Correlation: **Property Features are highly correlated -**



<----- Pearson Correlation heatmap of 43 property features



We have 3 different derivations of same feature as

1. median
2. mean
3. mode

These 3 derivations are more than 95% correlated with each other.

Selection: Keep features with high target correlation & less missing %

	corr_with_target %
floorsmax_avg	-4.30
floorsmax_medi	-4.28
floorsmax_mode	-4.19
floorsmin_avg	-3.51
floorsmin_medi	-3.48
floorsmin_mode	-3.32
elevators_avg	-3.26
livingarea_avg	-3.24
elevators_medi	-3.23
livingarea_medi	-3.22
totalarea_mode	-3.12
elevators_mode	-3.06
livingarea_mode	-2.97
apartments_avg	-2.91
apartments_medi	-2.91

Top 15 highly correlated property features with target

1. Of the 3 derivations, average is more correlated with target than mode and median. Hence, we drop mode & median.
We repeat the exercise for other highly correlated features.

3. Finally, **drop 38 features & just keep only the following features.**
   - *'basementarea_avg',*
   - *'floorsmax_avg',*
   - *'floorsmin_avg',*
   - *'landarea_avg',*
   - *'livingarea_avg'*

Similarly, we choose the features with less % of missing values from a highly correlated feature pair in correlations between
1. Non-Property continuous features
2. Categorical features
3. Continuous & Categorical features

Finally, we are left with **176 features**.

3 - Feature Engineering

Feature Engineering :

Based on EDA analysis, we create new features which help us in predicting credit default. These are different from the features we created while summarizing and merging the datasets in Data Wrangling stage. Couple of examples are provided below.

Application details based features -

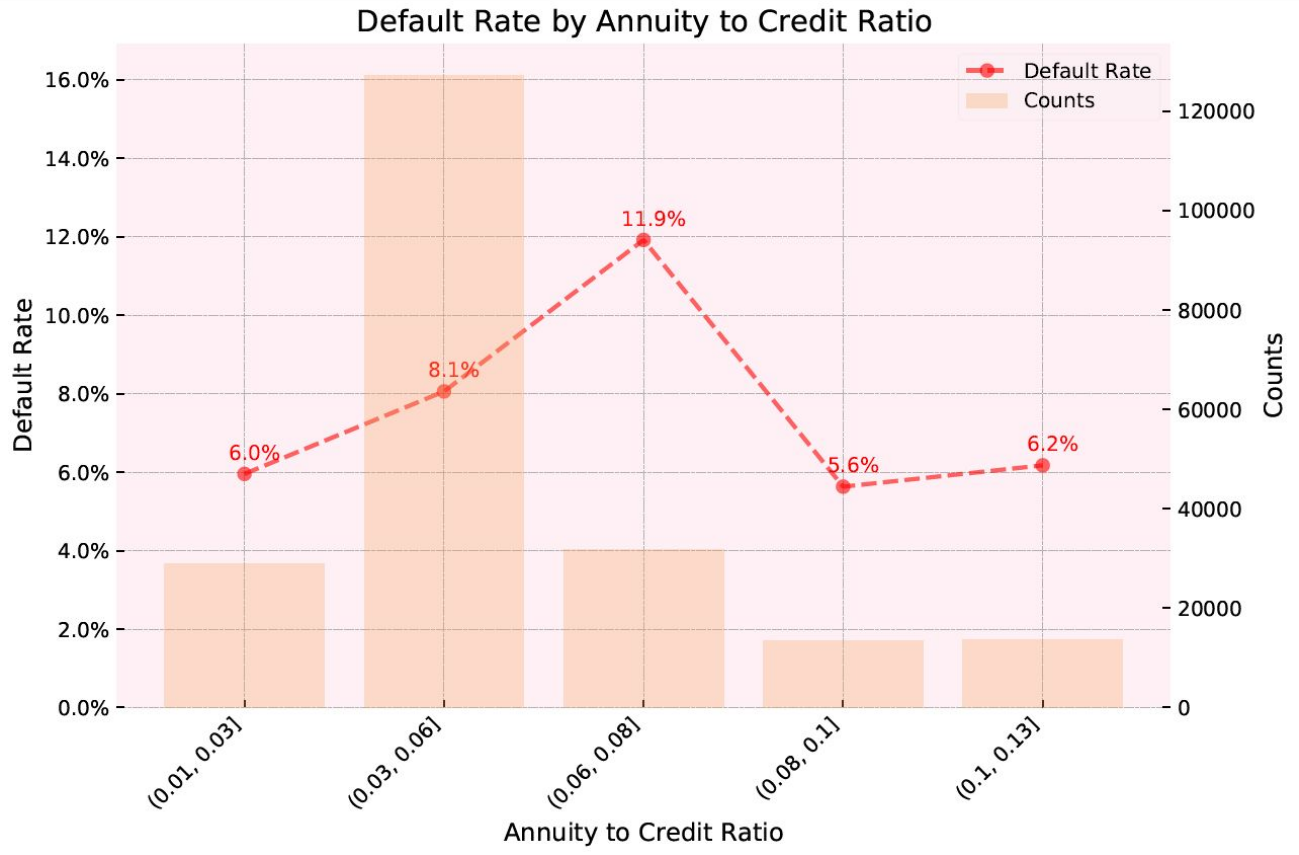
1. **rt_annuity_credit** - This is the ratio of required monthly annuity amount to be paid by customer to the credit amount.
2. **rt_goods_price_credit -** Ratio of goods price to credit provided.

Applicant details based features

1. **rt_days_employed_birth** - Ratio of number of days in current job to number of days since birth at the time of application
2. **avg_family_credit -** Credit amount per family member.

In next slides, we show the default rate trends in these features.

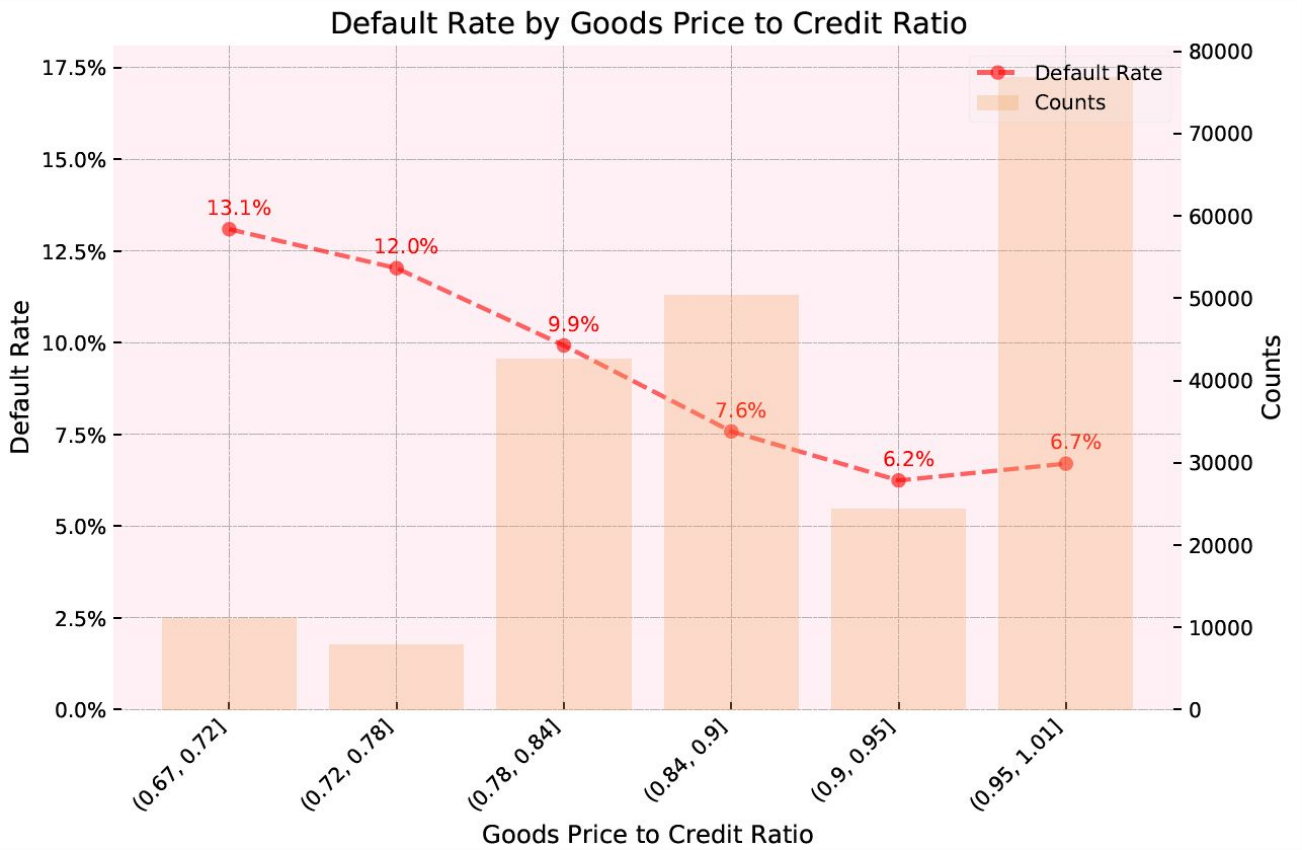
Annuity to Credit Amount ratio : **Lower default rate at tails**



This feature is a ratio of annuity to credit amount in current application.

Interestingly, we have lower default rate at tails and higher default rate in the middle.

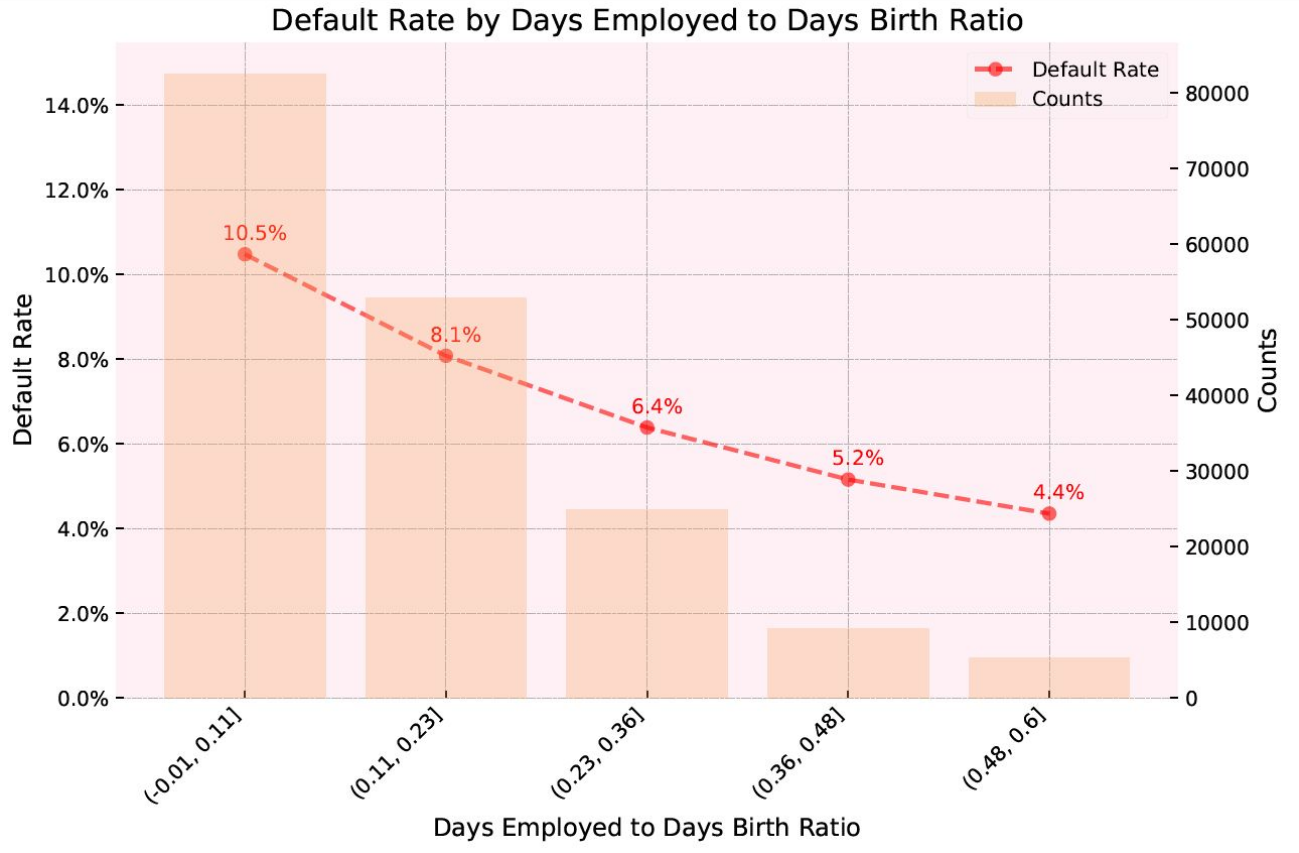
Goods Price to Credit Ratio : **Default rate decreases as the ratio increases**



Default rate decreases with increase in goods price to credit ratio.

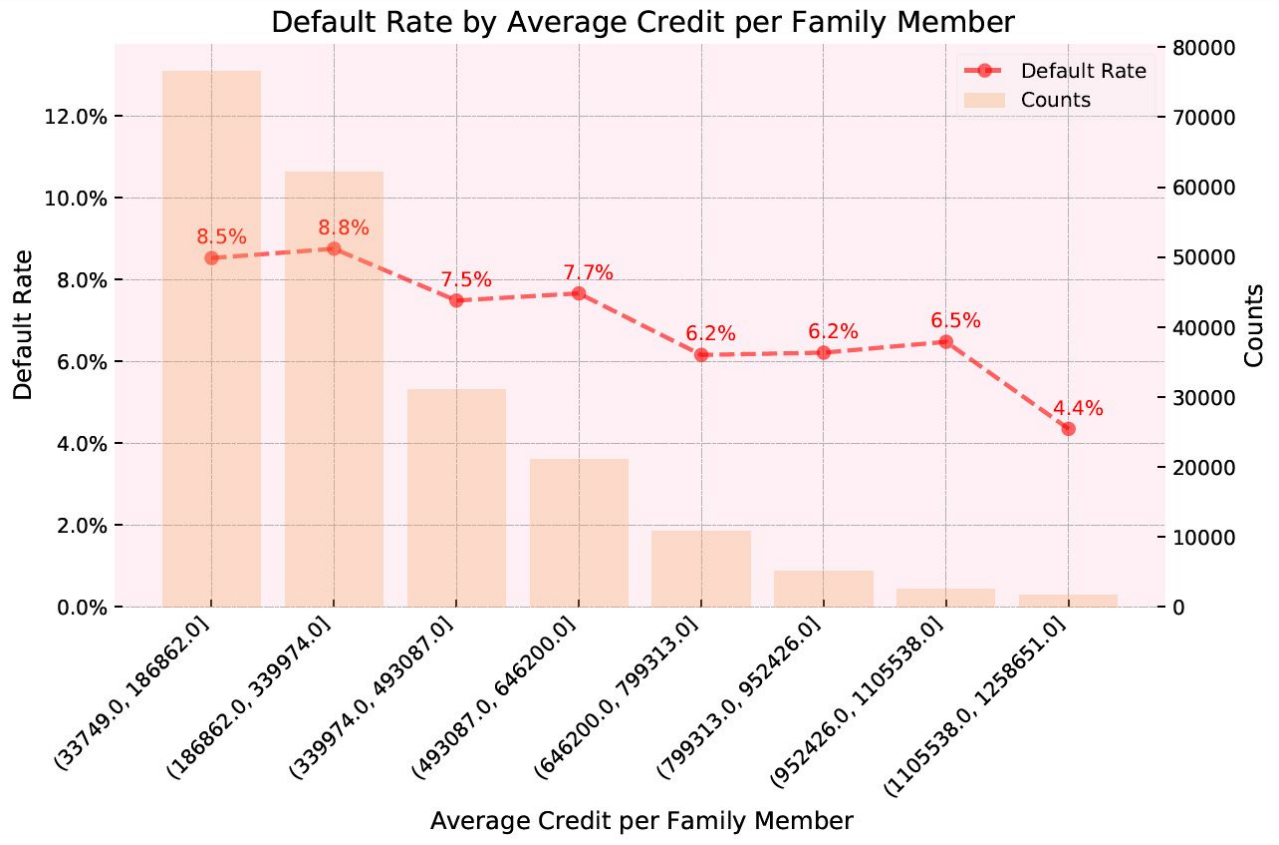
This may indicate that underwriting practices are strong where higher risks are given lower loan amounts (as % of goods price).

Days Employed to Days Birth :



Higher the number of days people are in their current job as a proportion of age in days, lower the default rate.

Credit per Family Member : Higher the credit per member, lower the default rate.



Higher the credit amount per family member, lower the default rate. Lower loan pay back capability may be because of higher expenses incurred in bigger families.

**It is possible to have multiple loans at the same time which is not being captured here.

Finally, we have **188 features** before the modelling stage.

4 - Modelling

Modelling - Evaluation & Models

1. Evaluation Metrics

We will evaluate the hyperparameter tuned models on below 3 matrices out of which **roc_auc** is the most important.
1. roc_auc score - For Home Credit, it is equally important to avoid customers who are likely to default on the loan and increase customers who are likely to pay back the loan. Hecne, we use the **roc_auc** score as it considers both the classes (default/no default). The pr_auc matrix mainly cares about positive class(default). Hence, we will avoid it. Higher the roc_auc score, better the model. We will tune threshold for profitability.
2. Prediction time - We will evaluate prediction time on train data after hyperparameter tuning. Lower the prediction time, better the model.
3. Tuning time - Lower the hyperparameter tuning time, better the model.

2. Models

HomeCredit is more interested in accurate predictions than interpretations.
1. **Logistic Regression with l1 regularization** - Logistic regression is a simple linear algorithm with decent performance. As we have 188 features and many are correlated, we will use **L1 regularization** for feature selection.
2. **XgBoost** - XgBoost is known to outperform linear models on tabular data in many cases.
3. **LightGBM** - LightGBM is faster than XgBoost and tends to perform well on tabular data.

Modelling - Missing Values & Validation

3. Handling Missing Values

1. **Categorical** features - Create a **seperate nan category**/level for missing value. We will do this at the time of one-hot-encoding using pandas get_dummy function.
2. **Numerical** features - **For logistic regression**, we will **impute missing values with median** because numerical features are highly skewed as observed in EDA. **XgBoost & LightGBM** can handle missing in a supervised way which optimizes model performance. Hence, we **do not impute** numerical features for these two algorithms.

4. Validation

We will use a **5-fold cross-validation** strategy on train data for model tuning and model selection. We will tune these 3 models on the same set train-validation datasets. To select the best hyperparameters, we will use **RandomSearchCV** technique. And we will compare the performance of the best model of each of the three algorithms. Best of the bests will be our final model.

And finally, we will test the performance of the final model on the test dataset. This will give us a better idea of how well the model performs on unseen data.

Modelling - Final Model

Best Model from RandomSearchCV	Iterations	Mean roc_auc score	Standard deviation of roc_auc score	Prediction Time*	Tuning Time
Logistic Regression	10	0.7643	0.0039	NA**	8min 35s
XgBoost	**50**	**0.7826**	**0.0047**	**20.4s**	**14h 6min 10s**
Light GBM	50	0.7814	0.0049	36.8s	1h 50min 49s

* Prediction time is calculated on the train dataset.
** As roc_auc score of logistic regression was not within 3 standard deviations of XgBoost model, prediction time was not evaluated.

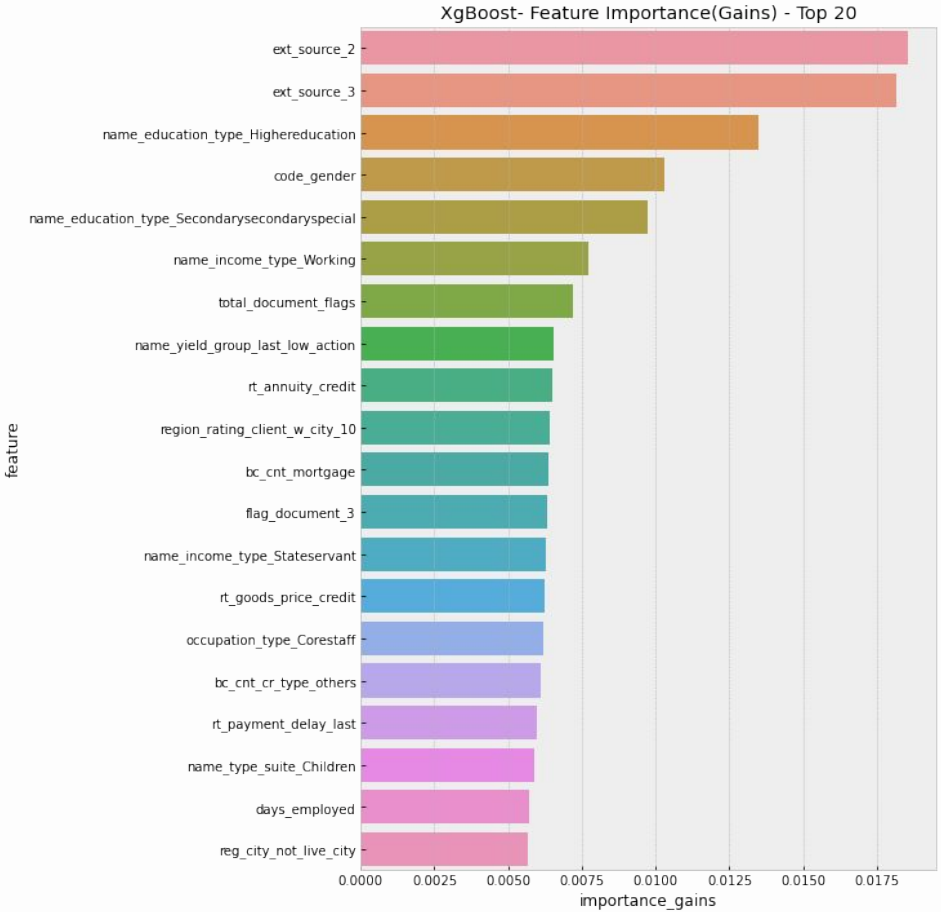
XgBoost is slightly better than Light GBM but roc_auc scores of both the models are within 1 standard deviation of each other. XgBoost prediction is 1.8 times faster than LightGBM prediction. LightGBM with 5 CPU cores is 7 times faster than XgBoost with 6GB GPU. **Hence, we pick XgBoost as our final model.**

```
xgb_final_params = {'learning_rate': 0.02332 ,  'max_depth' : 8,  'min_child_weight' : 1,
                    'colsample_bytree' : 0.6,  'subsample' : 0.8,  'n_estimators' : 800,
                    'reg_alpha' : 8.6410 }
```

Performance on Test Data

We get an AUC of **0.7841** on test data which is slightly higher than **0.7814** mean AUC of cross-validation but well within the range (0.7861, 0.7767) of 1 standard deviation. **The XgBoost model is well generalized on the test set.**

Modelling - Feature Importance



External feature 2 & 3 are the top two features. In EDA, we had observed that these features were exhibiting a strong trend with excellent risk differentiation (from 24% to 2%).

Demographic features like education, income & gender are important predictors of default rate.

Interestingly, 3 feature engineered ratio factors are among the top 20. These are annuity to credit, goods price to credit and payment delay to duration for last loan.

We also have 2 features from bureau data. They are counts of mortgage and others credit types.

We observed trends in many of the top 20 features during EDA.

5 - Profit Maximization

Assumptions & Approach

We assume that HomeCredit wants to maximize profit as a business objective. Also we make simplistic assumptions about below business parameters.
1. **5% of credit as a final profit** on loans which are fully paid - this is the average % profit after removing commission, expenses, cost of capital etc.
2. **40% of credit as a loss** on loans which are defaulted - this is the average % loss after considering outstanding amount, recoveries etc. 90% of loans are cash loans and losses tend to be higher on cash loans.

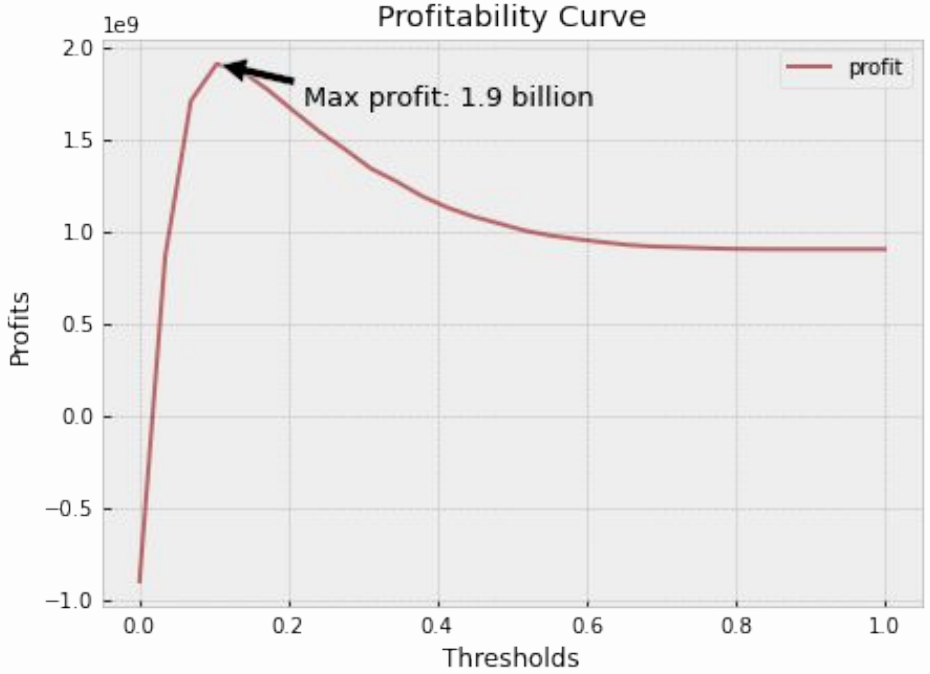
Expectancy =
+ tn * mean_credit_paid * profit_on_paid_loans (profit from fully repayable loans)
+ tp * mean_credit_default * loss_on_default_loans (notional profit from avoiding defaults)
- fp * mean_credit_paid * profit_on_paid_loans (notional loss from avoiding repayable loans)
- fn * mean_credit_default * loss_on_default_loans (loss from defaults on loans)

Average loan size for a paid loan (mean_credit_paid) = 601969.81.
Average loan size for a default loan (mean_credit_default) = 554235.12
**Average loan sizes are derived from test data.

Profitability Curve

1. Profitability Curve

We simulate various probability thresholds for classification and calculate expectancy on test data(30%).



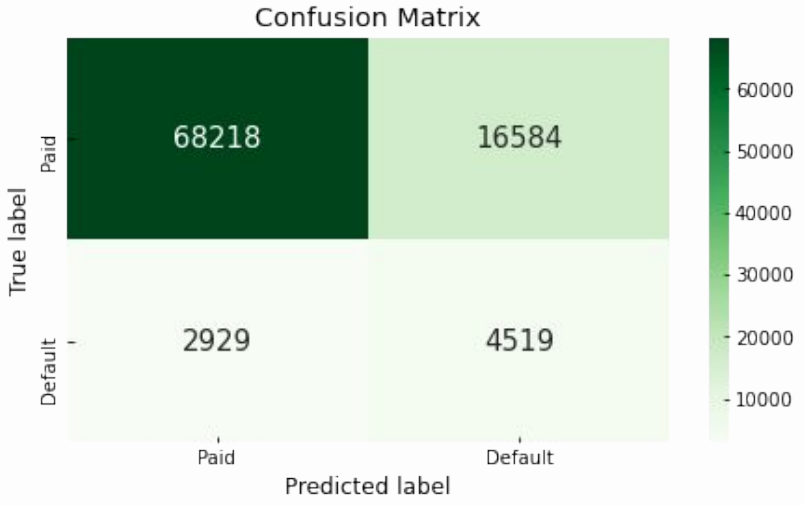
We get **0.1034** as profit maximizing threshold with **1.9 Billion** as maximum profit.

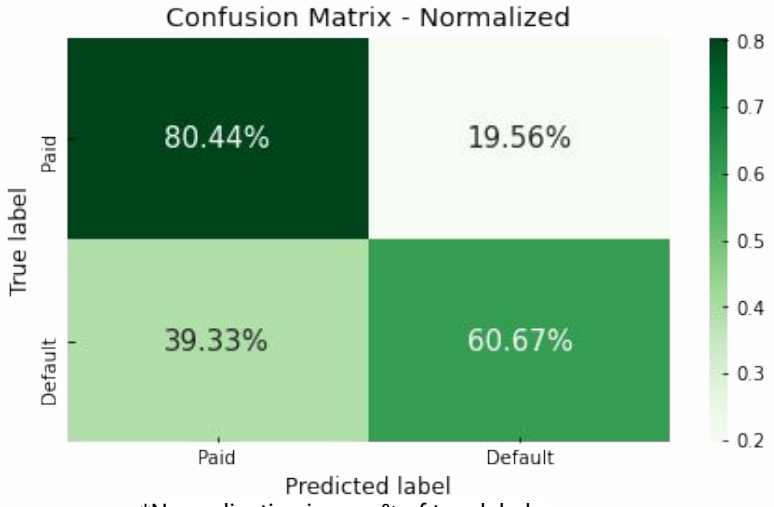
With 0.5 threshold, profit is 1 Billion which is ~45% of the best achievable profit.

Book Profit : ⬆ to 3.29% from 1.49%

2. Confusion Matrix

The confusion matrix for the profit maximizing threshold is





*Normalization is as a % of true labels.

With a 0.1034 threshold, we are likely to provide loans to **71147 (77.12%)** applications. We give loans to 80.44% of loan paying applications and accommodate 39.33% of the default applications. This essentially translated into a **book profit of 3.29%** (as a % of total credit extended). The current book profit is 1.49%

Further Improvements

Following ideas can be explored in the future

1. **Modelling applicants with and without credit card separately** - We have just about 28% applicants who had credit cards with HomeCredit but the records are credible(65k). This leads to missing values in most of the records. Also, applicants without credit cards have a 1.2% higher default rate. Separating these two groups may lead to better accuracy.
2. **Bayesian Hyperparameter Optimization** - which is a targeted search strategy based on model improvements. This will help in faster tuning and maybe in better performance.
3. **Model Explainability** - HomeCredit is more focused on prediction. With model Interpretation/Explainability analysis, we can demonstrate the impact of the individual feature for easy buy-in of the model by business.

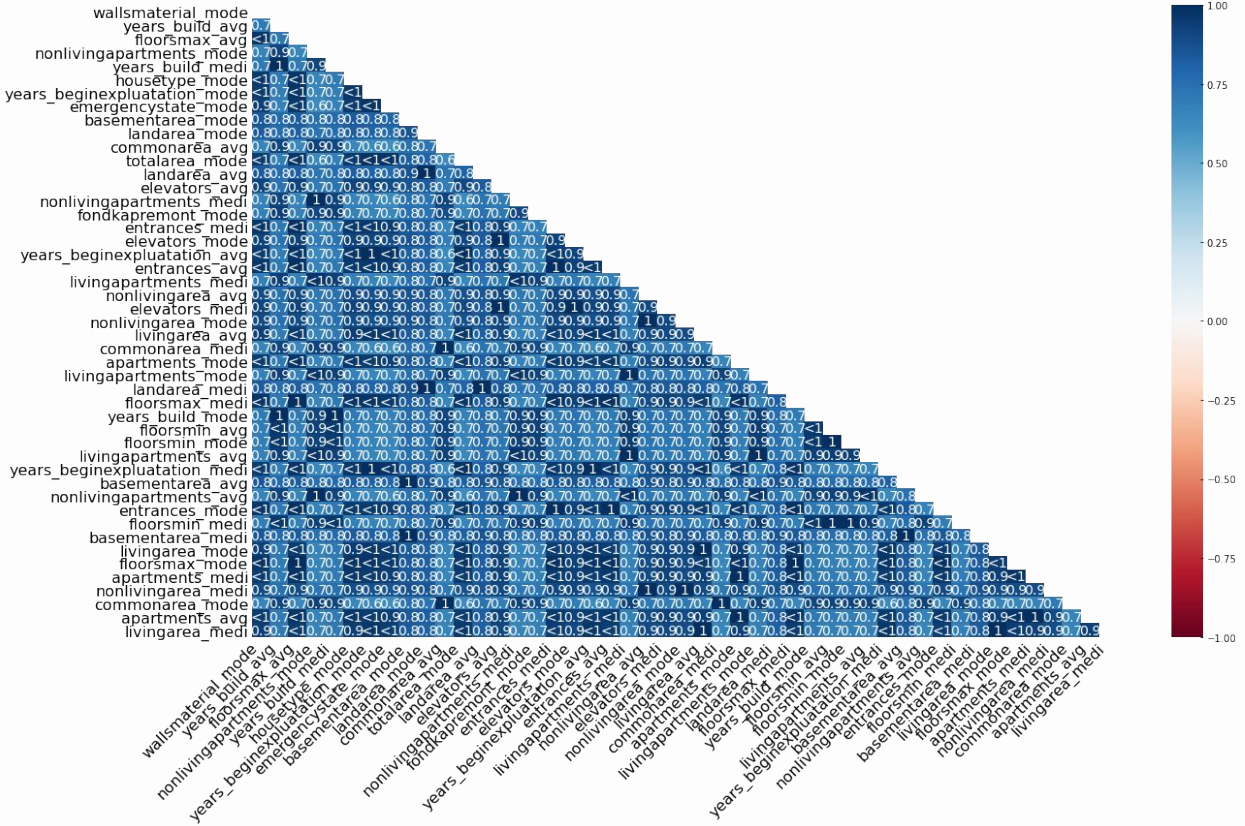
Appendix

1.1 Data Wrangling : Findings

#	Variables	Dataset	Findings/**Actions**
1	amt_annuity	application	Missing in 12 records.. It is an important feature. Hence, **delete 12 records**
2	days_employed	application	# of days since current employment. It is **365243** days(~1000 years) for pensioners. **Hence, replaced with NaN.**
3	days_first_due, days_last_due, days_termination etc	prev_application	For some records,, **+365243** days i.e. ~1000 years in the future value is present. Reasons are clients returned the goods soon after taking the loan, revolving loans etc. We **replace this odd value 365243 with nan.**
4	amt_credit > 1.5* amt_goods_price	application	Insurance cost is added to amt_credit in some cases. Assuming maximum insurance costs is 50% of a goods price, we still have 2564 records(<1%) failing the check.
5	amt_balance	credit_card_balance	Amount balance is negative in 0.06% cases. These customers overpaid what they own. Hence, we get a negative amount.
6	days_XX	all	All the dates are relative to the time of application and are negative.
7	code_gender	application	**In gender, 4 'XNA' values changed to 'F' (Female)**

Details can be found at Data Wrangling notebook.

1.2 Completeness : Property features have high missing %

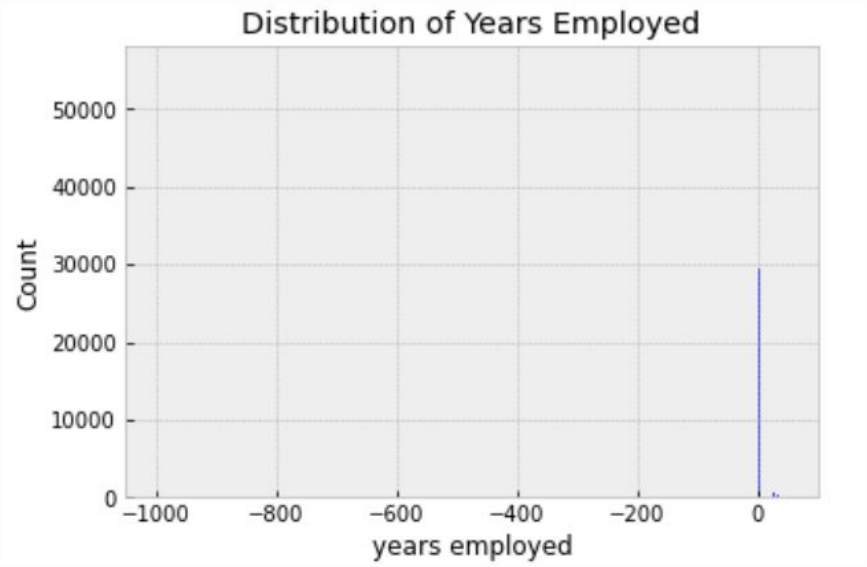


<---------- Heatmap of missing values correlation

We have 47 property related features with 47% to 69% missing values. And > 70% correlation indicates that the values are largely missing for same records.

Details can be found at Data Wrangling notebook.

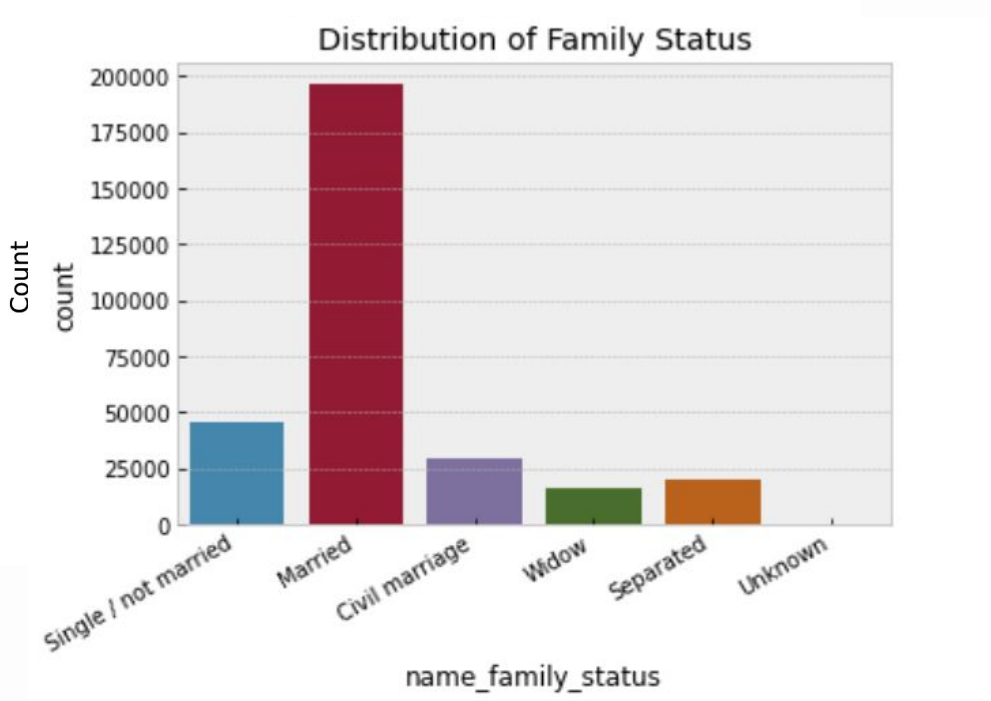
1.3 Validity of Years Employed



-1000 years does not seem to be a valid value for years. Closer inspection reveals that days employed is coded as 365243 for pensioners. After excluding pensioners, the distribution looks valid and reasonable.
**Raw feature is days employed. It is wrt application date & negative number. We divide it by -365.25 to convert it to years.

Details can be found at Data Wrangling notebook.

1.4 Accuracy of Family Status



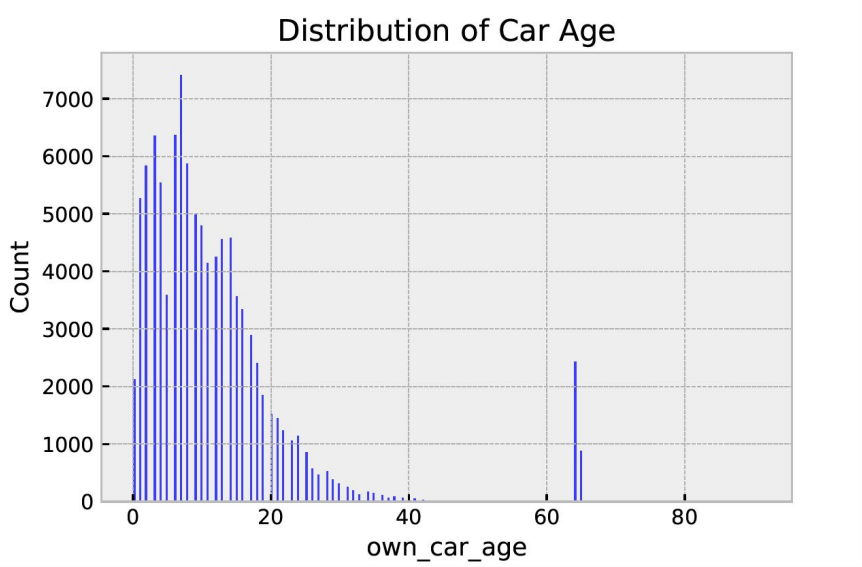
Family status has two values

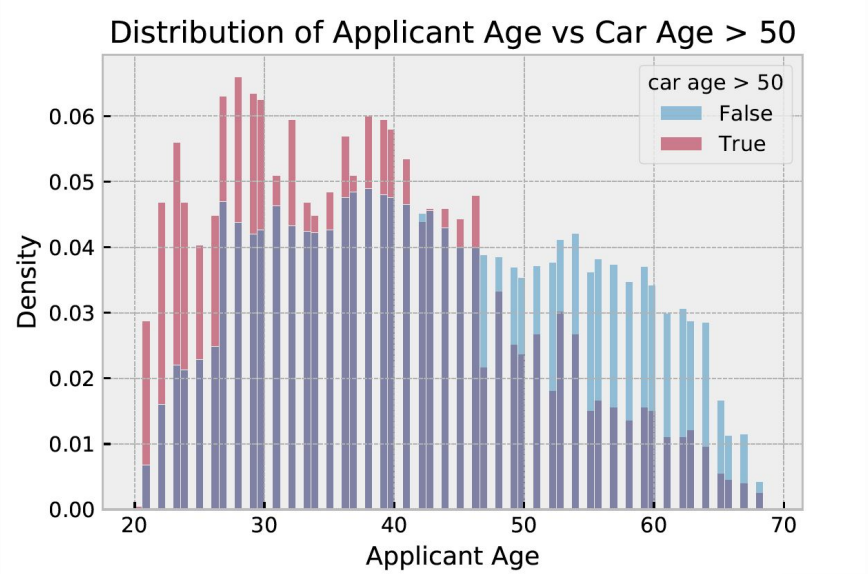
1. Married
2. Civil Marriage

It is worth checking the **accuracy** of these two categories with business. Whether all clients were presented these two options and/or they were aware of the difference between two options or there is way to verify this field.

Details can be found at Data Wrangling notebook.

1.5 Reasonableness of Car Age





We have a good number of cars with age 64-65. This may be a data aberration. Is it a verified field or declared field ? Is it that accidently people entered their own age instead of car's age ?

Proportion of young drivers is higher than proportion of old drivers for '> 50 years old' cars. We do not have any insight on why car age is higher based on above graph.

Details can be found at Data Wrangling notebook.

1.6 Combining the datasets

1

1. Create **behavioral & financial features for each transaction** eg
   Late Payments in days, Deficit in payments,
   Ratio of max credit balance to max credit limit etc

2. **Summarize loan transactions** to reflect mean, sum, maximum, most frequent and last attributes of all transactions of the previous loan.

3. Merge with Previous Application.

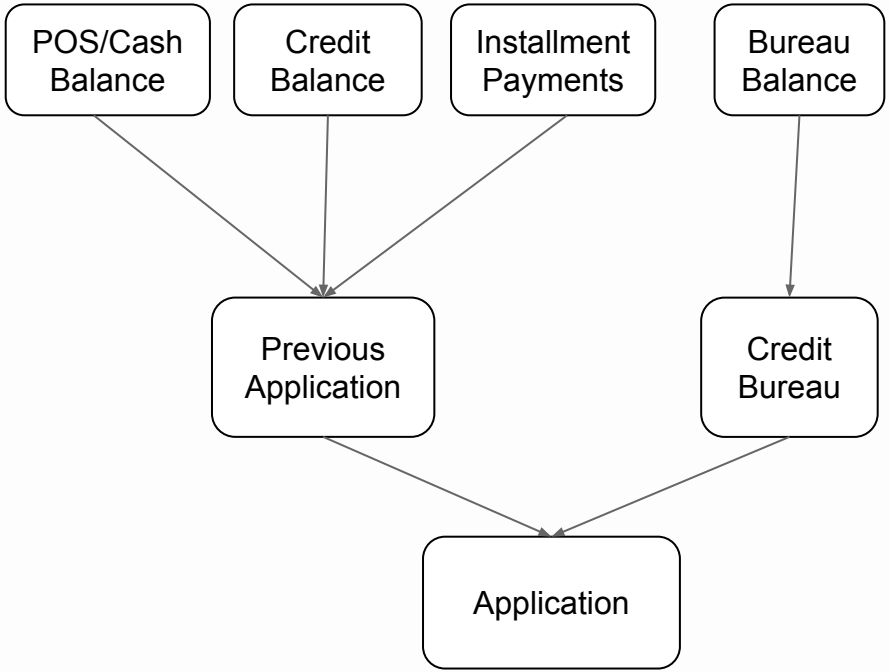
2

1. Create additional behavioral & financial features for **each previous loan** eg
   Ratio of annuity to credit amount,
   Ratio of total delay days to loan duration etc

2. **Summarize previous loans** using mean, maximum, most frequent and last attributes on previous loans.
   Last attributes reflects recent financial health.

3. Merge with Application data.

Repeat the same process for Bureau data.



307,499 records and 230 variables in combined data.

2.1 Exploratory Data Analysis (EDA) Approach

Train/Test Split -

We split the combined data into train and test sets. Train set is 70% with 215249 records and **Test set is 30%** with 92250 records. Using stratification, we ensured that both test and test sets have the same ie **8.07 % default loans**. **The EDA analysis is based on the train dataset.** This avoids leaking some information into test dataset.

Categorical features -

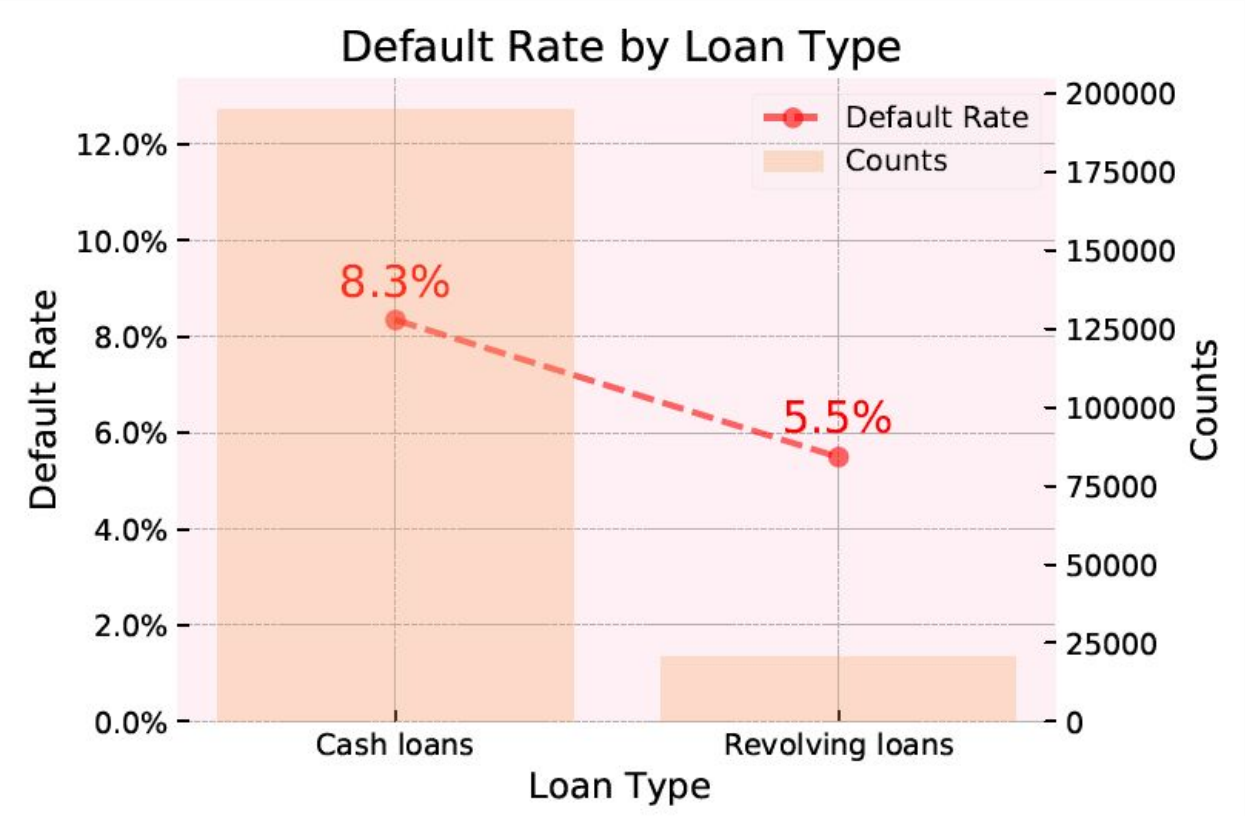
We will plot the feature on x-axis and default rate on primary y-axis. We will also plot the total number of observations (counts) in each category on the secondary y-axis. With counts, we can gauge the credibility of the default rate for that category. Missing value will be presented as separate Missing category.

Continuous features -

We will first bin the feature into a number of groups. This will smoothen out the trend and we will get a better picture when the trend is strong. We may remove 1 percentile records at top or bottom end in case of extreme values for the smoothness and clarity. This is just for visualization purposes and we are going to use the original continuous feature for modelling. Now, we plot bins on x-axis and default rate on primary y-axis. We will also plot the total number of observations (counts) in each bin on the secondary y-axis. We will exclude any missing data..

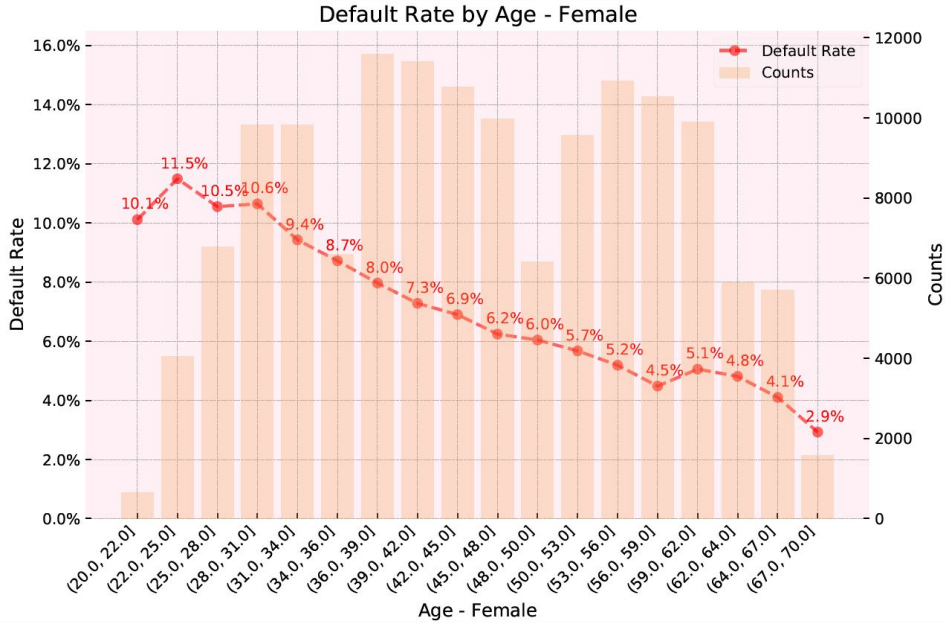
With this framework in mind, let us see what loan portfolio we have.

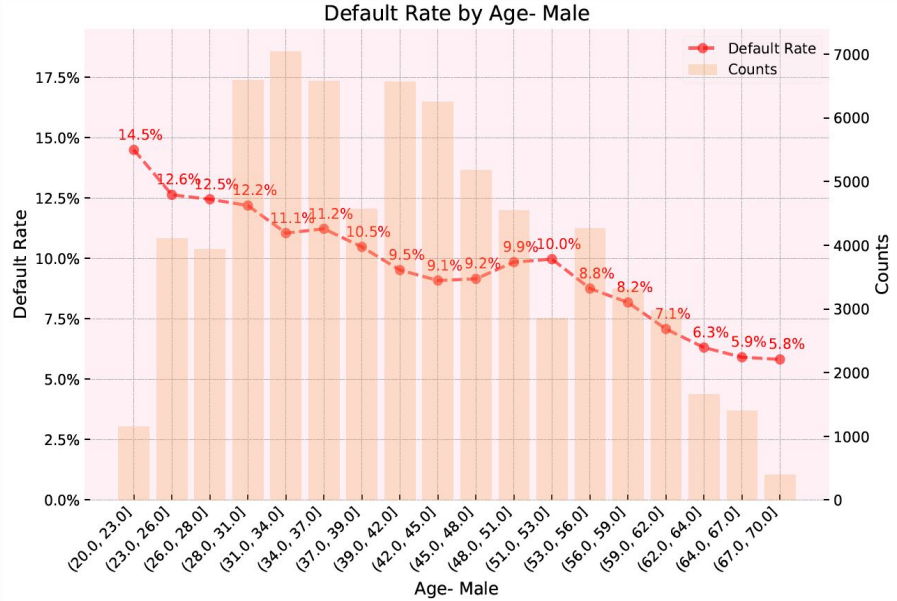
2.2 Loan Type : **90% loans are cash loans**



90% loans are cash loans in the data provided are cash loans. And cash loans have a higher default rate than revolving loans.

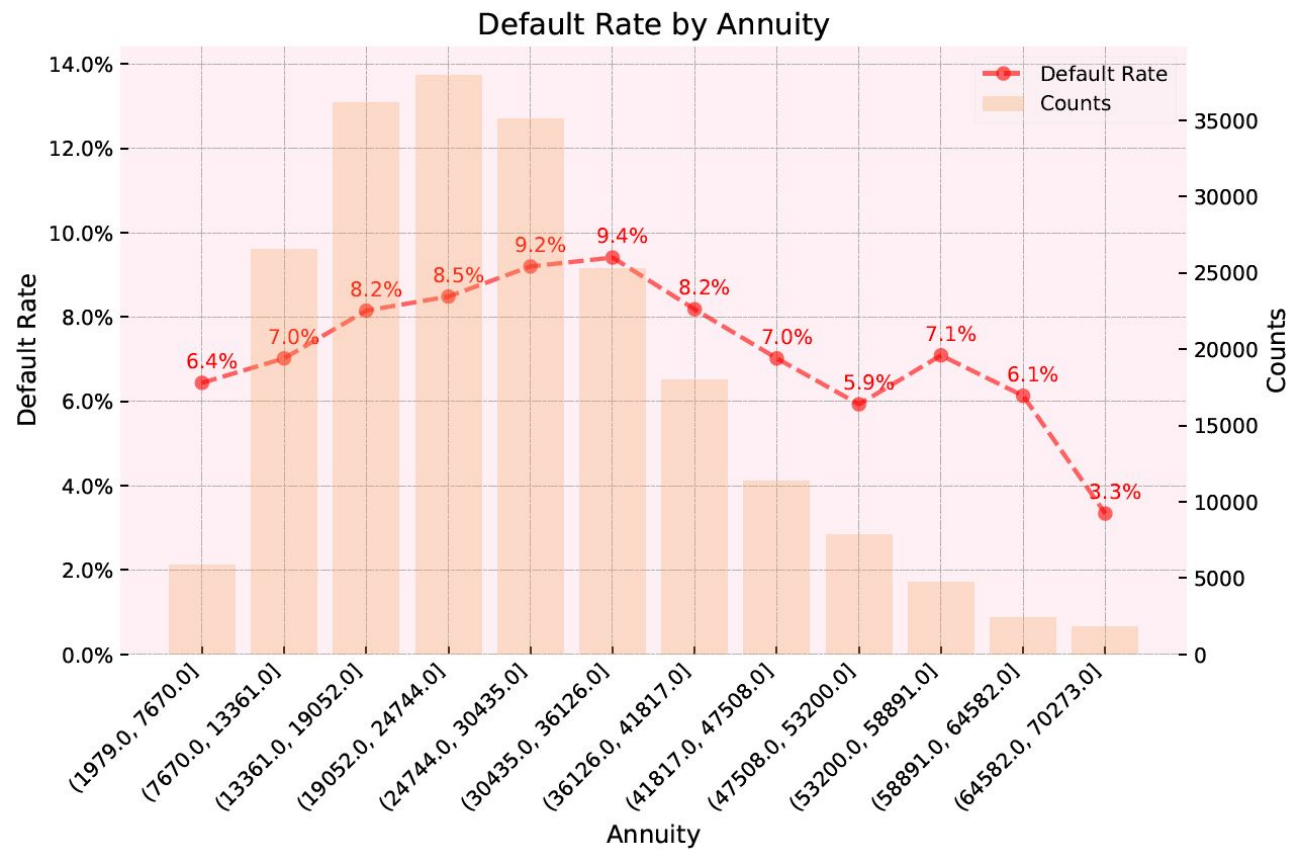
2.3 Age X Gender : **Decrease in default rate in middle ages is higher for females**





Just as we had observed for gender feature, the default rate is lower for females than for males. Apart from that, from age 28 to 56, default rate steeply decreases as female age increases. But, the decrease in default rate is slower for males in the same age range. This is indicating an interaction between age and gender features.

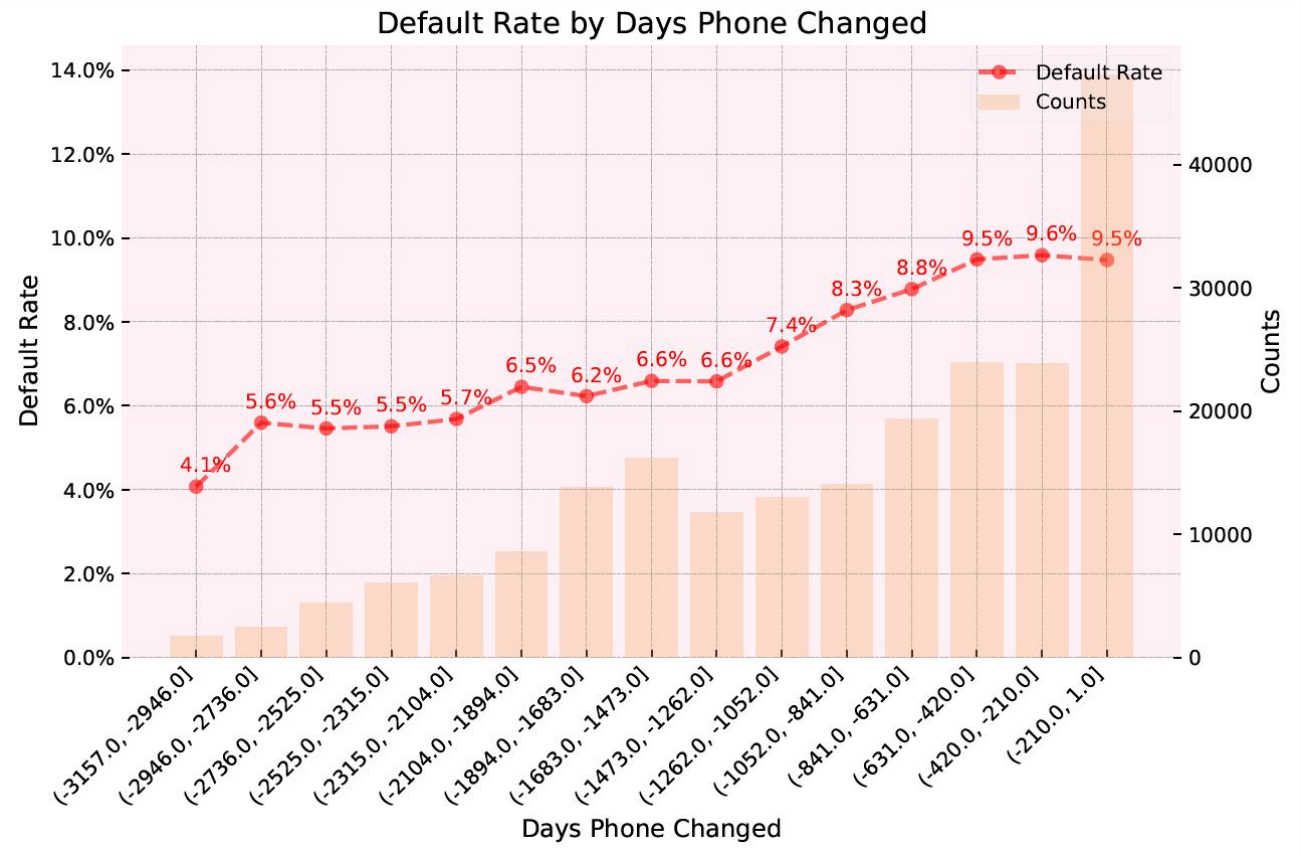
2.4 Annuity Amount : **Default rate has inverted U shape**



In Annuity amount, default rate is higher in the middle and then decreases as we move away from middle.

** Removed top 1 percentile for visualization purpose

2.5 Days Phone Changed : **The older the phone, lower the default rate**

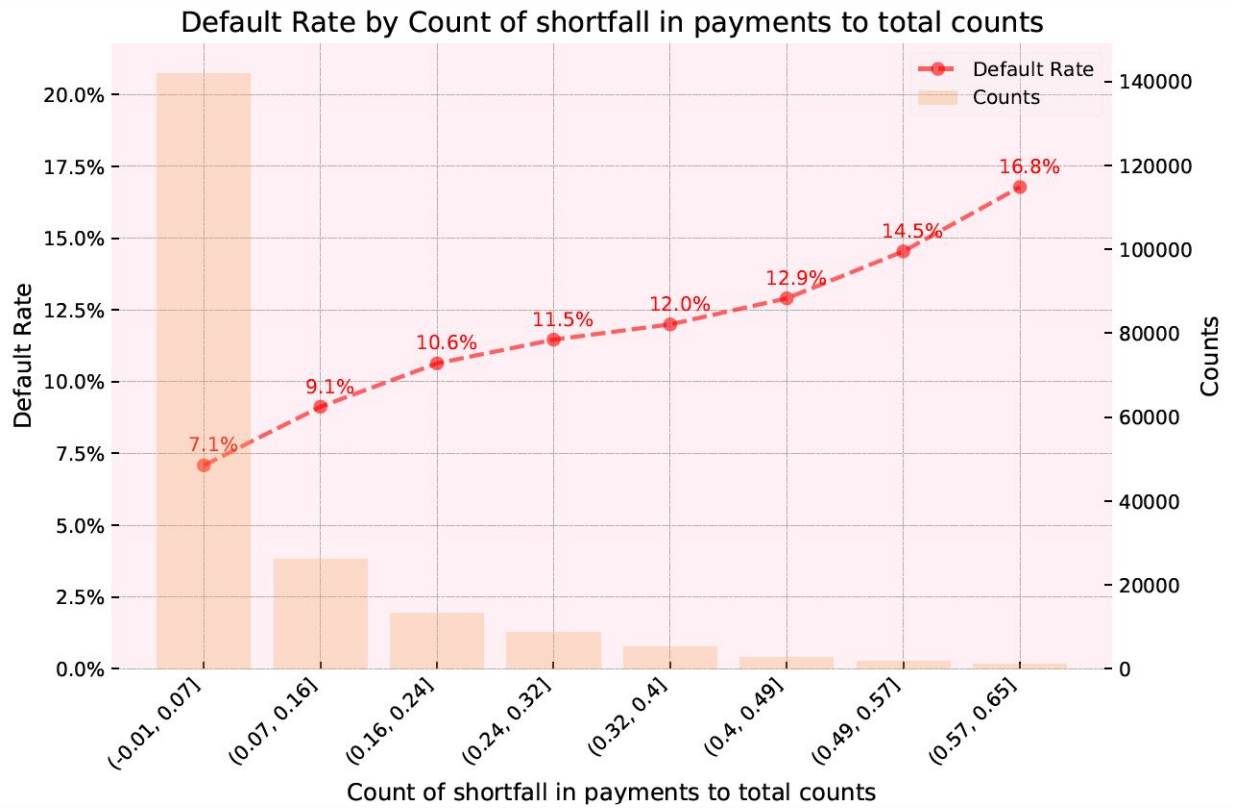


<---------- Older

The number of days before the application the client changed the phone. Days are wrt application date and are negative.

On average, the older the phone, the lower the default rate. This feature may be correlated with age.

2.6 Deficit Count on Past Loans - Normalized



The number of times there was a deficit in installment payment in past loans. It is normalized by total installment counts.

55% applications have paid required minimum installments amounts or more, indicated by 1st bin where all observed values are zeros. But as the deficit count increases, the default rate increases. This is intuitive as someone who is unable to pay the minimum installment amount is more likely to default.

** ~6% applications do not have past history and are missing and we are excluding top 1 percentile observations.

2.7 Correlation - Non Property(>90%) : **Keep one with less missing % of the pair**

feature 1	feature 2	absolute corr
cr_sum_cnt_drawings_curr_sum	cr_sum_cnt_drawings_curr_mean	99.92
cr_max_sk_dpd_def_max	cr_max_sk_dpd_def_last	99.90
obs_30_cnt_social_circle	obs_60_cnt_social_circle	99.84
cr_max_amt_cr_limit_last	cr_max_amt_cr_limit_max	99.83
amt_goods_price	amt_credit	98.70
bc_amt_max_credit_overdue	bb_last_loan_status	98.61
amt_credit_max	amt_goods_price_max	98.57
in_rt_cnt_deficit_pmt_mean	in_rt_amt_deficit_inst_mean	98.50
in_rt_cnt_deficit_pmt_max	in_rt_amt_deficit_inst_max	97.88
bc_cnt_loans	bc_cnt_consumer_credit	93.20
in_sum_payment_delay_mean	in_sum_payment_delay_last	92.89
bc_cnt_closed	bc_cnt_loans	92.35
amt_goods_price_mean	amt_credit_mean	91.99
bc_cnt_closed	bc_cnt_consumer_credit	91.73

Highly correlated features are:

1. *Derivations of same feature* eg sum & mean, mean & last etc
2. *Derivations of same pair of features* eg (amt_goods_price_max , amt_credit_max), (amt_goods_price_mean , amt_credit_mean)

It makes sense to drop one feature from each pair. And **we will drop the feature which has more missing values within the pair.** But we will keep both amt_goods_price & amt_credit in raw form.

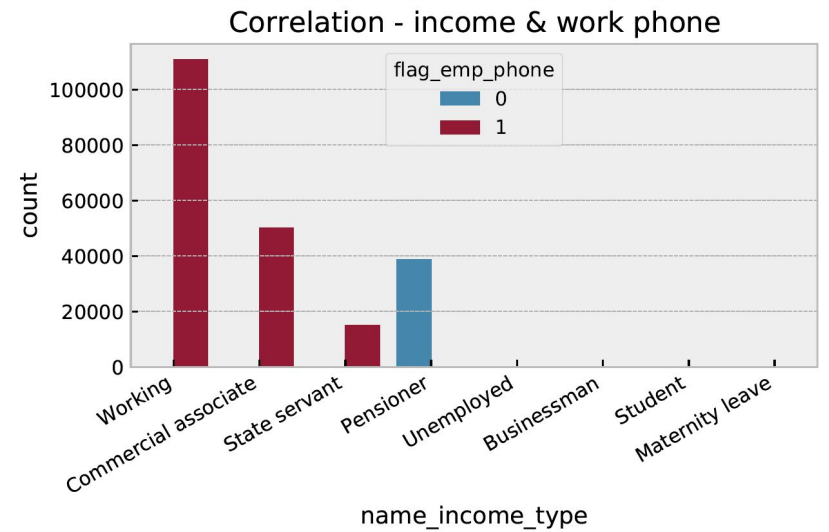
Dropped features :

- *'cr_sum_cnt_drawings_curr_mean', 'amt_goods_price_mean',*
- *'Cr_max_sk_dpd_def_max', 'obs_60_cnt_social_circle',*
- *'Cr_max_amt_cr_limit_max', 'amt_goods_price_max',*
- *'In_rt_cnt_deficit_pmt_mean', 'in_rt_cnt_deficit_pmt_max',*
- *'Bc_cnt_loans', 'in_sum_payment_delay_mean',*
- *'Bc_cnt_loans', 'bc_cnt_consumer_credit'*

**** Property and Non-Property features are not highly correlated with other.**

2.8 Correlation : **Categorical Features**

	feature_1	feature_2	cramers_v
309	name_income_type	flag_emp_phone	99.98
599	flag_emp_phone	organization_type	99.98
888	region_rating_client	region_rating_client_w_city	95.73
1070	reg_region_not_work_region	live_region_not_work_region	86.18
1973	name_yield_group_last	product_combination_last	82.96



We have 64 categorical features. The table contains the top 5 correlation pairs.

Income vs Employee Phone:
From the figure, it is clear that pensioners did not provide employer provided work phone numbers in credit applications. Similar logic applies to organization type.

Region rating & region rating with the city are providing almost the same information.

Hence, we **drop the employer provided phone flag & region rating features.**

We do not see a strong correlation between continuous & categorical features. Hence, it is not included in the presentation

3.1 Feature Engineering :

Application details based features -

1. **rt_annuity_credit** - This is the ratio of required monthly annuity amount to be paid by customer to the credit amount.
2. **rt_goods_price_credit -** Ratio of goods price to credit provided.
3. **rt_credit_income -** Ratio of credit to income. Income is not a verified field but a declared field.
4. **rt_annuity_income -** Ratio of annuity to income. This indicates the annuity payment capability wrt to income.
5. **total_document_flags -** Sum of document provided indicators. This indicates completeness of application.

Applicant details based features

1. **rt_days_employed_birth** - Ratio of number of days in current job to number of days since birth at the time of application
2. **avg_family_credit -** Credit amount per family member.
3. **rt_days_id_birth -** Ratio of number of days in current job to number of days since last change in ID before loan application.
4. **rt_phone_changed_birth -** Ratio of number of days in current job to number of days since phone change ID before loan application.
5. **avg_family_income -** Income amount per family member.
6. **total_contact_flags -** Sum of contacts points provided indicator.

Process overview

1	Data Wrangling	1. Explore the data for completeness, validity, accuracy & reasonableness. 2. Combine all seven datasets into one where each application is one row and each feature is one column.
2	Exploratory Data Analysis	1. Split the data into train(70%) & test(30%) set. 2. Explore the distribution of features, their relationship with target variables. Identify features showing strong trends wrt target. 3. Explore relationship between features & select one from highly(>90%) correlated pair.
3	Feature Engineering	1. Based on EDA & business sense, we create new features. 2. Application based features are annuity to credit ratio, total documents submitted etc. 3. Applicant based features are credit per family member, years employed by age etc.
4	Modelling	1. Select appropriate evaluation matrix. 2. Use 5-fold RandomSearchCV for hyperparameter tuning and test data for validation. 3. Train logistic regression, XgBoost & LightGBM models. 4. Select the best model and validate on test data.
5	Business Impact	1. Select the profit maximization probability threshold for classifying loans. 2. Analyze the impact on profitability. 3. Finally, communicate the results using a presentation & a technical report.