Home Credit – Credit Risk Analysis

**Team: Group\_ML1**

* Jiaoyan Zhang
* Aruna Vedula
* Vijay Bhaskaran
* Varun Malhotra

2018

Table of Contents

[1. Introduction 2](#_Toc522136623)

[2. Literature Review 2](#_Toc522136624)

[3. Data 3](#_Toc522136625)

[4. Exploratory Data Analysis 6](#_Toc522136626)

[4.1 Examine the Distribution of the Target Column 6](#_Toc522136627)

[4.2 Examine Missing Values 6](#_Toc522136628)

[4.3 Column Type 7](#_Toc522136629)

[4.4 Aligning Training and Testing Data 7](#_Toc522136630)

[4.5 Anomalies 7](#_Toc522136631)

[4.6 Correlations 9](#_Toc522136632)

[4.7 Effect of Age on Repayment 10](#_Toc522136633)

[4.8 Exterior Sources 13](#_Toc522136634)

[4.9 Pairs Plot 16](#_Toc522136635)

[5. Model Building and Analysis 18](#_Toc522136636)

[6. Conclusion 20](#_Toc522136637)

[Reference: 22](#_Toc522136638)

[Appendix: 23](#_Toc522136639)

# 1. Introduction

Risk Analysis is a proven way of identifying and assessing factors that could negatively affect the success of a business or project. It allows you to examine the risks that you or your organization face and helps you decide whether to move forward with a decision. You do a Risk Analysis by identify threats and estimating the likelihood of those threats being realized. Once you have worked out the value of the risks you face, you can start looking at ways to manage them effectively. This may include choosing to avoid the risk, sharing it, or accepting it while reducing its impact.

It's essential that you're thorough when you're working through your Risk Analysis, and that you're aware of all the possible impacts of the risks revealed. This includes being mindful of costs, ethics, and people's safety.

The objective of this competition is to use historical loan application data to predict whether an applicant will be able to repay a loan. This is a standard supervised classification task:

* Supervised: The labels are included in the training data and the goal is to train a model to learn to predict the labels from the features
* Classification: The label is a binary variable, 0 (will repay the loan on time), 1 (will have difficulty repaying the loan)

Home Credit Group is a leading international multi-channel provider of consumer finance founded in the Czech Republic in 1997, with operations on 3 continents. A service dedicated to providing lines of credit (loans) to the unbanked population. The challenge is to find out how capable each applicant of home credit is of repaying a loan? Poor or non-existent credit histories create a challenge to obtain loans. Hence, non – regularized lenders take advantage of the above situation. The idea here is clients with repayment abilities should never be rejected Home Credit believes in fostering positive and safe borrowing experience among the applicants. In times, where every decision is based on the data and Home Credit considers this project a critical business need. While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data.

Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities.

# 2. Literature Review

Based on Edward and other researchers’ (1992) work, they intended to find the determinant factors of default risk for house loans. They conducted multivariate logit models and explored the relevance of payment history, loan terms, borrower characteristics, economic conditions, and legal constraints in order to analyze the loan defaults and delinquencies. Payment history emerges as the overwhelming factor in predicting the default probabilities.

In 2010, Amir and other scientists also conduct research on consumer credit risk based on machine learning modeling algorithms. They combined credit bureau data and client transactions in the past five years sampling customers from a giant commercial bank, they constructed out-of-sample forecasts to greatly improve the classification rates of credit card holder defaults and delinquencies and the linear regression R-squared of forecasted/realized delinquencies of 85%.

Henri in 2005 found that financial contracts may be redesigned to allow for banks to manage the idiosyncratic factor for their accounts while allowing separately dealing with the systematic cause. The systematic risk can be retained, shared with the borrower, or passed to the capital markets.

# 3. Data

The data is provided by Home Credit (<http://www.homecredit.net/about-us.aspx>), a service dedicated to provided lines of credit (loans) to the unbanked population. Predicting whether a client will repay a loan or have difficulty is a critical business need, and Home Credit is hosting this competition on Kaggle to see what sort of models the machine learning community can develop to help them in this task.

There are 7 different sources of data:

application\_train/application\_test: the main training and testing data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK\_ID\_CURR. The training application data comes with the TARGET indicating 0: the loan was repaid or 1: the loan was not repaid.

bureau: data concerning the client's previous credits from other financial institutions. Each previous credit has its own row in bureau, but one loan in the application data can have multiple previous credits.

bureau\_balance: monthly data about the previous credits in the bureau. Each row is one month of a previous credit, and a single previous credit can have multiple rows, one for each month of the credit length.

previous\_application: previous applications for loans at Home Credit of clients who have loans in the application data. Each current loan in the application data can have multiple previous loans. Each previous application has one row and is identified by the feature SK\_ID\_PREV.

POS\_CASH\_BALANCE: monthly data about the previous point of sale or cash loans clients have had with Home Credit. Each row is one month of a previous point of sale or cash loan, and a single previous loan can have many rows.

credit\_card\_balance: monthly data about previous credit cards clients have had with Home Credit. Each row is one month of a credit card balance, and a single credit card can have many rows.

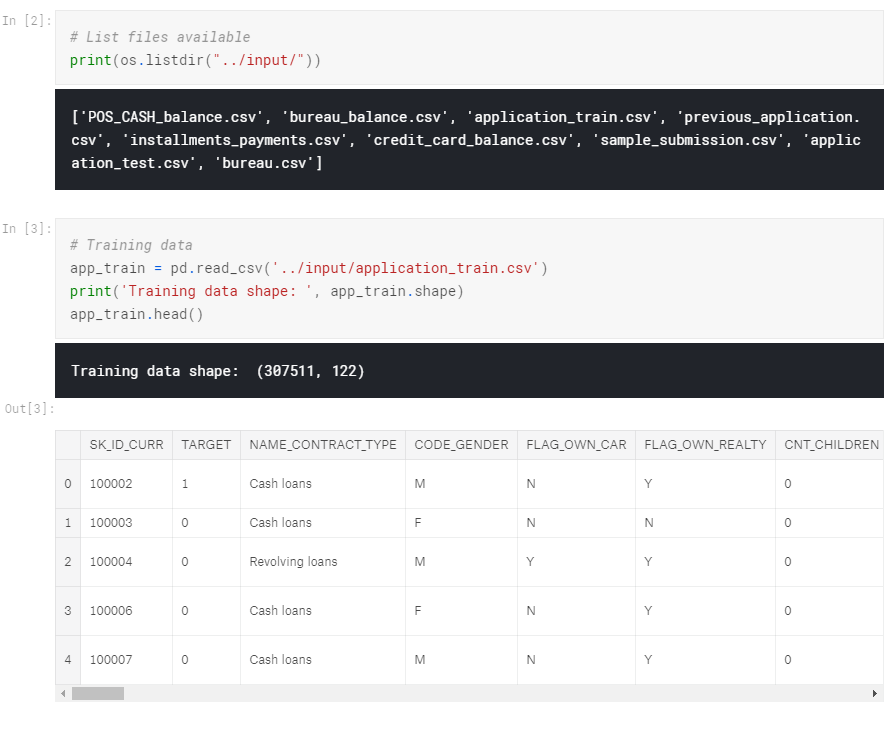
installments\_payment: payment history for previous loans at Home Credit. There is one row for every made payment and one row for every missed payment.

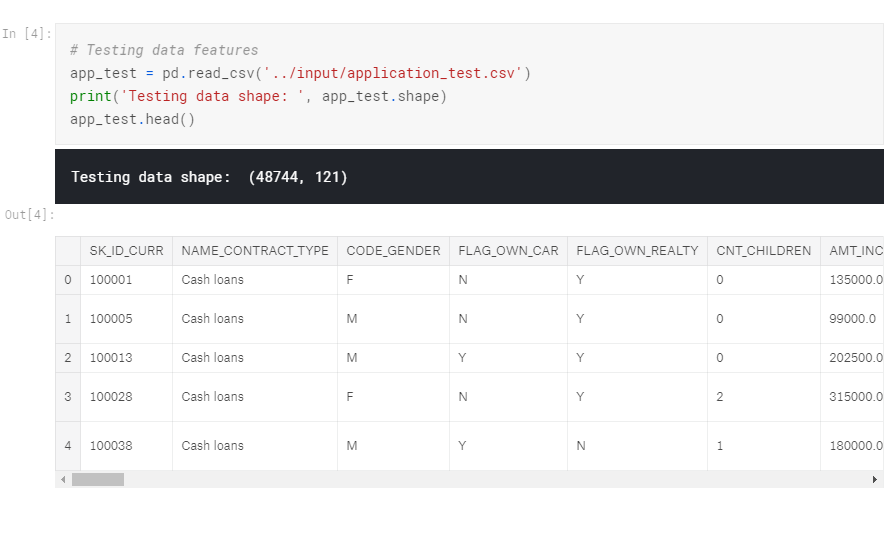
The diagram below shows how the data is related:



First, we can list all the available data files. There are a total of 9 files: 1 main file for training (with the target) 1 main file for testing (without the target), 1 example submission file, and 6 other files containing additional information about each loan.

The training data has 307511 observations (each one a separate loan) and 122 features (variables) including the TARGET(the label we want to predict).



The test set is considerably smaller and lacks a TARGET column.

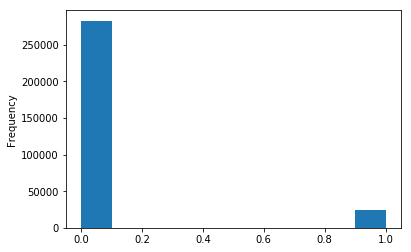
# 4. Exploratory Data Analysis

In this section, we calculate statistics and make figures to find trends, anomalies, patterns, and any relationships within the data set. The goal of EDA is to learn what our data can tell us. We will start with a high-level overview, then narrows in to specific areas as we find intriguing areas of the data. We will utilize our findings to inform the next modeling choices, such as by helping us decide which features to use.

## 4.1 Examine the Distribution of the Target Column

The target is what we are asked to predict: either a 0 for the loan was repaid on time, or a 1 indicating the client had payment difficulties. We can first examine the number of loans falling into each category.

We see this is an imbalanced class based on the chart above. There are far fewer loans that were not repaid on time than loans that were. To reflect the imbalance, we can weight the classes by the data representation if we conduct more sophisticated machine learning models.



## 4.2 Examine Missing Values

Next, we can look at the number and percentage of missing values in each column. Among the 122 columns, there are 67 columns with missing values.

When it comes time to build our machine learning models, we will have to fill in these missing values. We will use models such as XGBoost that can handle missing values with no need for imputation. Another option would be to drop columns with a high percentage of missing values, although it is impossible to know ahead of time if these columns will be helpful to our model. Therefore, we will keep all of the columns for now.

## 4.3 Column Type

Let's look at the number of columns of each data type. int64 and float64 are numeric variables (which can be either discrete or continuous). object columns contain strings and are categorical features.

Most of the categorical variables have a relatively small number of unique entries. We will need to find a way to deal with these categorical variables. Let's implement the policy described above: for any categorical variable (dtype == object) with 2 unique categories, we will use label encoding, and for any categorical variable with more than 2 unique categories, we will use one-hot encoding. For label encoding, we use the Scikit-Learn Label Encoder and for one-hot encoding, the pandas get\_dummies(df) function.

## 4.4 Aligning Training and Testing Data

There need to be the same features (columns) in both the training and testing data. One-hot encoding has created more columns in the training data because there were some categorical variables with categories not represented in the testing data. To remove the columns in the training data that are not in the testing data, we need to align the data frames. First, we extract the target column from the training data because this is not in the testing data but we need to keep this information. When we do the align, we set axis = 1 to align the data frames based on the columns and not on the rows.

The training and testing datasets now have the same features which are required for machine learning. The number of features has grown significantly due to one-hot encoding. At some point, we probably will want to try dimensionality reduction (removing features that are not relevant) to reduce the size of the datasets.

## 4.5 Anomalies

One problem we always want to be on the lookout for when doing EDA is anomalies within the data. These may be due to mis-typed numbers, errors in measuring equipment or they could be valid but extreme measurements. One way to support anomalies quantitatively is by looking at the statistics of a column using the described method. The numbers in the DAYS\_BIRTH column are negative because they are recorded relative to the current loan application. To see these stats in years, we can mutliple by -1 and divide by the number of days in a year.

count 307511.000000

mean 43.936973

std 11.956133

min 20.517808

25% 34.008219

50% 43.150685

75% 53.923288

max 69.120548

Name: DAYS\_BIRTH, dtype: float64

Those ages look reasonable. There are no outliers for the age on either the high or low end. Then we will take a look at the days of employment.

count 307511.000000

mean 63815.045904

std 141275.766519

min -17912.000000

25% -2760.000000

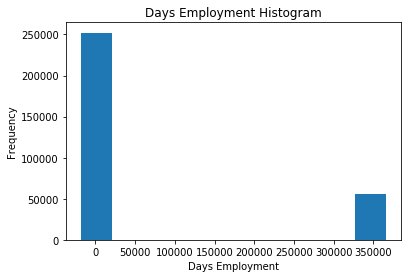
50% -1213.000000

75% -289.000000

max 365243.000000

Name: DAYS\_EMPLOYED, dtype: float64

As the maximum value (besides being positive) is about 1000 years, we would say that the above numbers are not good.



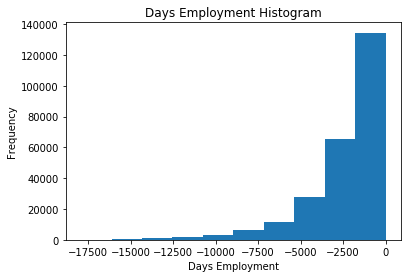
The non-anomalies default on 8.66% of loans

The anomalies default on 5.40% of loans

There are 55374 anomalous days of employment

After we subset the anomalous clients, as we can see they tend to have low rates of default than the rest of the clients.

Handling the anomalies depends on the exact situation, with no set rules. One of the safest ways is to set the anomalies to a missing value and then have them filled in via Imputation before conduct machine learning. In this case, since all the anomalies have the same value, we want to fill them in with the same value in case all of these loans share something in common. The anomalous values seem to have some importance, so we want to tell the machine learning model if we did, in fact, fill in these values. As a solution, we will fill in the anomalous values with not a number (np.nan) and then create a new boolean column indicating whether or not the value was anomalous.



The distribution looks to be much more in line with what we would expect, and we also have created a new column to tell the model that these values were originally anomalous. The other columns with DAYS in the dataframe look to be about what we expect with no obvious outliers. Similarly, we will apply the same method to the testing data. There are 9274 anomalies in the test data out of 48744 entries.

## 4.6 Correlations

Next, we will look for correlations between the target and features. We will calculate the Pearson correlation coefficient between every variable and the target, which indicates an idea of possible relationships within the data.

Most Positive Correlations:

OCCUPATION\_TYPE\_Laborers 0.043019

FLAG\_DOCUMENT\_3 0.044346

REG\_CITY\_NOT\_LIVE\_CITY 0.044395

FLAG\_EMP\_PHONE 0.045982

NAME\_EDUCATION\_TYPE\_Secondary / secondary special 0.049824

REG\_CITY\_NOT\_WORK\_CITY 0.050994

DAYS\_ID\_PUBLISH 0.051457

CODE\_GENDER\_M 0.054713

DAYS\_LAST\_PHONE\_CHANGE 0.055218

NAME\_INCOME\_TYPE\_Working 0.057481

REGION\_RATING\_CLIENT 0.058899

REGION\_RATING\_CLIENT\_W\_CITY 0.060893

DAYS\_EMPLOYED 0.074958

DAYS\_BIRTH 0.078239

TARGET 1.000000

Name: TARGET, dtype: float64

Most Negative Correlations:

EXT\_SOURCE\_3 -0.178919

EXT\_SOURCE\_2 -0.160472

EXT\_SOURCE\_1 -0.155317

NAME\_EDUCATION\_TYPE\_Higher education -0.056593

CODE\_GENDER\_F -0.054704

NAME\_INCOME\_TYPE\_Pensioner -0.046209

DAYS\_EMPLOYED\_ANOM -0.045987

ORGANIZATION\_TYPE\_XNA -0.045987

FLOORSMAX\_AVG -0.044003

FLOORSMAX\_MEDI -0.043768

FLOORSMAX\_MODE -0.043226

EMERGENCYSTATE\_MODE\_No -0.042201

HOUSETYPE\_MODE\_block of flats -0.040594

AMT\_GOODS\_PRICE -0.039645

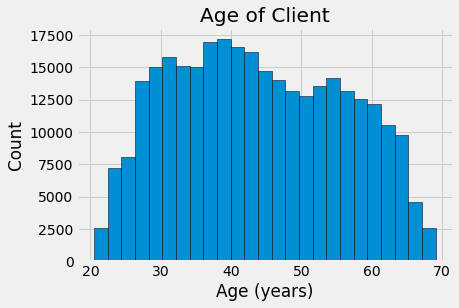
REGION\_POPULATION\_RELATIVE -0.037227

Name: TARGET, dtype: float64

Based on the results, the DAYS\_BIRTH is the most positive correlation. DAYS\_BIRTH is the age in days of the client at the time of the loan in negative days. The correlation is positive, but the value of this feature is actually negative, meaning that as the client gets older, they are less likely to default on their loan. We decide to take the absolute value of the feature and then the correlation would be negative.

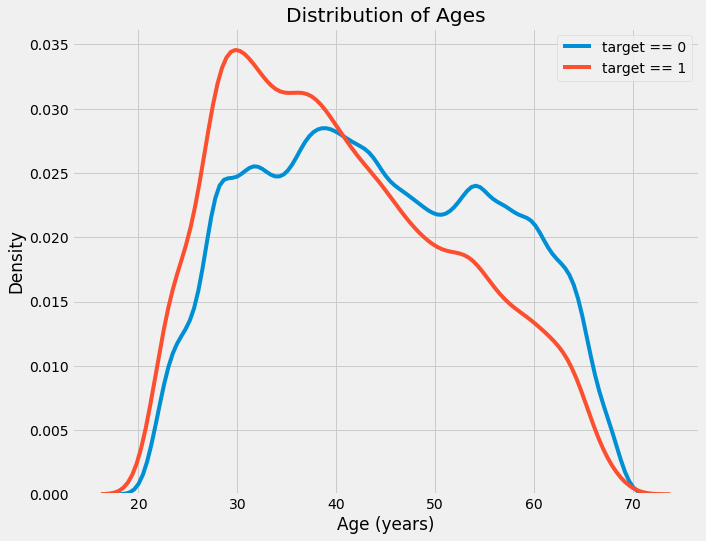
## 4.7 Effect of Age on Repayment

We find the correlation between the positive days since birth and the target is -0.078. This indicates that as the client gets older, there is a negative linear relationship with the target meaning that as clients get older, they tend to repay their loans on time more often. Firstly, we plot a histogram of the client age in years.



The distribution of age shows no outliers as the ages are all pretty reasonable. To visualize the effect of the age on the target, we will then plot a kernel density estimation plot (KDE) colored by the value of the target.

The kernel density estimate plot shows the distribution of a single variable as a smoothed histogram. We will use the seaborn kdeplot for this graph.

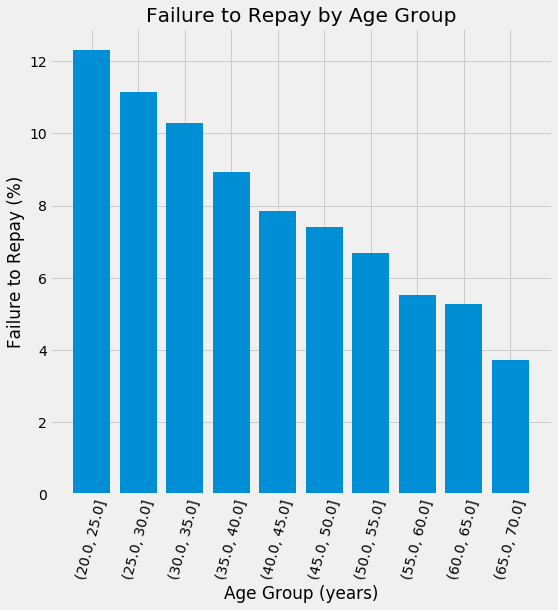


The target == 1 curve skews towards the younger end of the range. Although this is not a significant correlation (-0.07 correlation coefficient), this variable is likely going to be useful in a machine learning model because it does affect the target.

Let's look at this relationship in another way: average failure to repay loans by age bracket. To make this graph, first, we cut the age category into bins of 5 years each. Then, for each bin, we calculate the average value of the target, which tells us the ratio of loans that were not repaid in each age category.

There is a clear trend: younger applicants are more likely to not repay the loan. The rate of failure to repay is above 10% for the youngest three age groups and beolow 5% for the oldest age group.

This is information that could be directly used by the bank: because younger clients are less likely to repay the loan, maybe they should be provided with more guidance or financial planning tips. This does not mean the bank should discriminate against younger clients, but it would be smart to take precautionary measures to help younger clients pay on time.



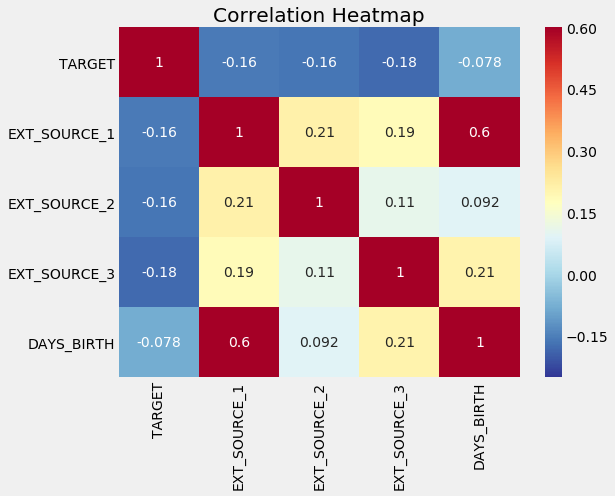
## 4.8 Exterior Sources

The 3 variables with the strongest negative correlations with the target are EXT\_SOURCE\_1, EXT\_SOURCE\_2, and EXT\_SOURCE\_3. According to the documentation, these features represent a "normalized score from external data source", which could be a cumulative sort of credit rating made using numerous sources of data.

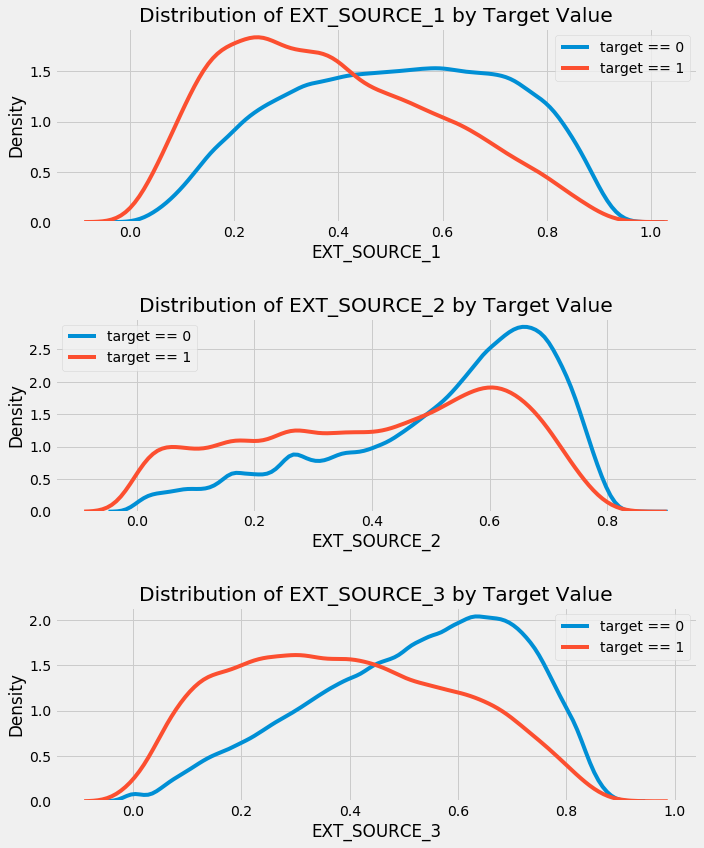
First, we can check the correlations of the EXT\_SOURCE features with the target and with each other.

|  | **TARGET** | **EXT\_SOURCE\_1** | **EXT\_SOURCE\_2** | **EXT\_SOURCE\_3** | **DAYS\_BIRTH** |
| --- | --- | --- | --- | --- | --- |
| **TARGET** | 1.000000 | -0.155317 | -0.160472 | -0.178919 | -0.078239 |
| **EXT\_SOURCE\_1** | -0.155317 | 1.000000 | 0.213982 | 0.186846 | 0.600610 |
| **EXT\_SOURCE\_2** | -0.160472 | 0.213982 | 1.000000 | 0.109167 | 0.091996 |
| **EXT\_SOURCE\_3** | -0.178919 | 0.186846 | 0.109167 | 1.000000 | 0.205478 |
| **DAYS\_BIRTH** | -0.078239 | 0.600610 | 0.091996 | 0.205478 | 1.000000 |

All three EXT\_SOURCE features have negative correlations with the target, indicating that as the value of the EXT\_SOURCE increases, the client is more likely to repay the loan. We can also see that DAYS\_BIRTH is positively correlated with EXT\_SOURCE\_1 indicating that maybe one of the factors in this score is the client age.



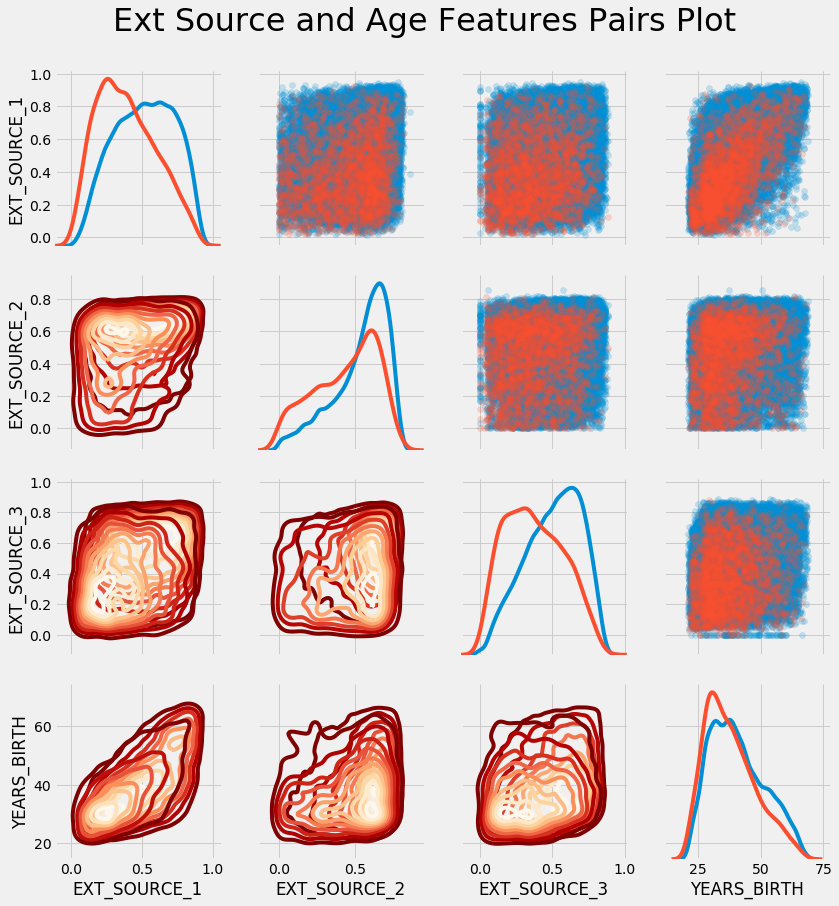
Next, we can look at the distribution of each of these features colored by the value of the target. This will let us visualize the effect of this variable on the target.



## 4.9 Pairs Plot

We will conduct a final pairs plot of the DAYS\_BIRTH and EXT\_SOURCE variables. This report uses the seaborn visualization library and the PairGrid function to create a Pairs Plot with scatterplots on the upper triangle, histograms on the diagonal, and 2D kernel density plots and correlation coefficients on the lower triangle.

The red in the plot means loans that were not repaid and the blue are loans that are paid. We could find the different relationships in the data. A moderate positive linear relationship exists between the the DAYS\_BIRTH and EXT\_SOURCE\_1, reflecting that the feature probably takes into account the client age.



# 5. Model Building and Analysis

After having performed the necessary data wrangling and EDA, it was time to perform feature Engineering. It is important to retain the crux of the data but reduce the no of variables for ease in model building. Looking at the correlations of the variables “EXT\_SOURCE\_1”, “EXT\_SOURCE\_2”, “EXT\_SOURCE\_3, we see that they are all negatively correlated with the target variable, which shows that probably they are similar in some way. Hence, we decided to employ the technique of feature engineering which helps in the process of choosing the right features or variables that are most significant in building our model. Feature engineering could be of two types: Feature construction and feature selection. Feature construction comprises of constructing or adding new features which comprise of the existing data and feature selection is the method of selecting only the most important features in the model building, in an attempt to reduce the “curse of dimentionality”. Here we employed one method of feature construction, namely Polynomial Features.

**Polynomial features**

Though two variables on their own might night have a great influence on the target, a new variable created by interacting the two variables might have a greater influence in the target and that is what we did with the three “EXT\_SOURCE “ variables and the DAYS\_BIRTH variable. We create polynomial features using the EXT\_SOURCE variables and the DAYS\_BIRTH variable and kept the degree to 3, as the features can scale exponentially with the degree, and can create problems in handling and overfitting. Now the 4 features have increased to 35 features which are either combination of two features or are original features raised to a power. The next step is to see how are the features correlated with the target variable. Finally, these new features were added to a copy of the test and train datasets so that, we could do analysis with and without using the polynomial features.

**Model Building**

The first model we build is the logistic regression model, which is a fairly simple, binary classification model, which is easy to build and has high interpretability, i.e., ease of understanding and interpreting the results. We use Scikit-Learn to build the model and use the model to make predictions. The logistic regression model was built on the dataset that did not use the polynomial features. We use the default model settings in building our logistic model using SciKit-Learn, except for one small change in the settings. We lower the regularization parameter, “c”, which will lower the overfitting in the model and thus will help us get a slightly better model with better prediction power. Since the prediction of our model is for a Kaggle competition, which asks for a submission of the predictions in probabilities, we use the “predict. proba” function to predict the probabilities of our submission. Upon submitting our predictions, our logistic model scored a prediction of 0.671

The next model that we created was the Random Forests model. Random forests is a very powerful classification and regression algorithm that is much like a decision tree, but unlike Decision trees, Random Forests does not only develop one tree. Instead, it grows hundreds of trees using the bagging technique, and hence, it is not prone to problems like overfitting like in decision trees. To make the model more efficient, we used the dataset with the polynomial features to see if the polynomial features boost the prediction power of the model. Again like with the logistic regression model, we find the probability of our predictions and submit it again to check if there is any change in the prediction power. The Random forests model with the polynomial features in the dataset scores 0.678 when submitted.

# 6. Conclusion

The objective of this project is to use the historical loan application data available and try to predict using Machine Learning technique if an applicant will be able to repay the loan or not. The labels are included in the data set and the goal is to train the model to learn and predict the label from the features. We used the Logistic Regression model, which is a binary classification model that identifies YES OR NO. We wanted to compare the accuracy of the Logistic regression model with Random Forest Model. Random Forest is a powerful classification and regression algorithm that is much like a decision tree, but it doesn’t grow just one tree but a bag of trees using the bagging technique. We used polynomial features to boost the prediction power of the model.  Analysis showed that the Random Forest technique had a higher accuracy percentage than Logistic Regression.

**Limitation:**

Although we would have used the important features for prediction, we lack the subject matter expertise in this analyze, that could have added much more value to the analysis. We did not cross validate the data, we used the data as is. The prediction is based on One data set and this cannot be used as the universal prediction. More data points of this kind can help us predict the accurate output from locations.

**Future Work:**

This project can be further improved with this future work. Perform feature selection by using domain knowledge, and feature importance. Use advanced techniques like gradient boosting, which is used for regression and classification problems, and it builds the model in a stage wise fashion, just like a decision tree or Xgboost to further enhance the model’s prediction.  Use better metrics for model evaluation like Area under ROC.

**Contributions:**

Team name on Kaggle: Group\_ML1

Scores for two submissions:

1. Logistic Regression: 0.604
2. Random Forests: 0.604

* ****Jiaoyan Zhang: researched on the literature review, conducted Exploratory Data Analysis (EDA analysis)
* Aruna Vedula: worked on the feature selection and model building
* Vijay Bhaskaran: conducted the data section
* Varun Malhotra

# Reference:

Amir E. Khandani, Adlar J. Kim, Andrew W. Lo, Consumer credit-risk models via machine-learning algorithms, Journal of Banking & Finance, Volume 34, Issue 11, 2010, Pages 2767-2787, ISSN 0378-4266, <https://doi.org/10.1016/j.jbankfin.2010.06.001>. (<http://www.sciencedirect.com/science/article/pii/S0378426610002372>)

Edward C. Lawrence, L.Douglas Smith, Malcolm Rhoades, An analysis of default risk in mobile home credit, Journal of Banking & Finance, Volume 16, Issue 2,1992, Pages 299-312, ISSN 0378-4266, <https://doi.org/10.1016/0378-4266(92)90016> S.(<http://www.sciencedirect.com/science/article/pii/037842669290016S>)

Henri Loubergé, Harris Schlesinger, (2005) "Coping with credit risk", The Journal of Risk Finance, Vol. 6 Issue: 2, pp.118-134, https://doi.org/10.1108/15265940510585798

# Appendix:

Below is the link to the Python code:

<https://drive.google.com/open?id=1Ks_Xbi50512n6-WUyJSRkNMSUz4uaXzV>