Project Proposal Group 7

Lisa Wang, Javier Cervantes, Rakeen Rouf, Kelly Tong

Dataset 1: Estimating a Company's Credit Risk

Overview:

This dataset has been fused from multiple sources:

- The bond list we got from the holdings of the USIG Ishares Credit Bond ETF. The dataset is a subset of the entire holdings such that our dataset is comprised only of companies in the **S&P500**: General Info Link to CSV
- Each company's fundamentals ratios were sourced from Yahoo Finance using the **yfinance** package
- Each company's credit rating (Fitch, Moody's, S&P) were sourced from **Bloomberg Terminal**
- The social sentiment indicators were sourced from **Finhubb API**

The entire dataset was sourced *Sept 22, 2023*

Research Questions:

Research Question 1 (Ordinal Outcome Variable):

- **Question**: Can we anticipate if a certain bond's credit rating will receive an upgrade/downgrade from the rating agencies?
- **Outcome Variable**: average_credit_rating

Research Question 2 (Continuous Outcome Variable):

- **Question**: Can we predict a certain bond's credit spread based on various metrics like the company's fundamentals and the market's sentiment related to that company?
- **Outcome Variable**: credit_spread

Dataset 1 Variable Description:

`Name:`

Source: ishares

Description: The bond issuer's name

`Sector:`

Source: ishares

Description: the issuer's main industry sector

`CUSIP:`

Source: ishares

Description: a bond's unique identifier

`ISIN:`

Source: ishares

Description: a bond's unique identifier

`Price:`

Source: ishares

Description: each bond's price

`Duration:`

Source: ishares

Description: measures the sensitivity of a bond´s price to changes in interest rates

`YTM (%):`

Source: ishares

Description: total return anticipated on a bond if the bond is held until it matures

`Maturity:`

Source: ishares

Description: the date when the bond's principal is repaid in full

`Coupon (%):`

Source: ishares

Description: each bond's yearly payment to investors as a percent of the bond's principal

ticker:

Source: ishares

Description: the issuer's stock ticker

marketCapitalization:

Source: ishares

Description: the value of the issuer's publicly traded stock

shareOutstanding:

Source: ishares

Description: the number of the issuer's shares outstanding

mention:

Source: Finhubb API (end point - **/stock/social-sentiment?symbol=ticker)**

Description: Number of mentions on Reddit in the last 1 year

negativeMention:

Source: Finhubb API (end point - **/stock/social-sentiment?symbol=ticker)**

Description: Number of negative mentions on Reddit in the last 1 year.

positiveMention:

Source: Finhubb API (end point - **/stock/social-sentiment?symbol=ticker)**

Description: Number of positive mentions on Reddit in the last 1 year.

negativeScore:

Source: Finhubb API (end point - **/stock/social-sentiment?symbol=ticker)**

Description: Sum of hourly negative scores (range 0 to 1) over the last 1 year

positiveScore:

Source: Finhubb API (end point - **/stock/social-sentiment?symbol=ticker)**

Description: Sum of hourly positive scores (range 0 to 1) over the last 1 year

score:

Source: Finhubb API (end point - **/stock/social-sentiment?symbol=ticker)**

Description: Sum of hourly net scores (range -1 to 1) over the last 1 year

RTG_FITCH:

Source: Bloomberg

Description: each bond's long term credit rating assigned by Fitch

RTG_SP:

Source: Bloomberg

Description: each bond's long term credit rating assigned by S&P

operating_profit_margin:

Source: yfinance

Description: operating profit / total revenue

ebitda_margin:

Source: yfinance

Description: EBITDA / total revenue

roa:

Source: yfinance

Description: return on assets = net income / total assets

debt_to_assets:

Source: yfinance

Description: total debt / total assets

debt_to_equity:

Source: yfinance

Description: total debt / stockholder's equity

int_coverage:

Source: yfinance

Description: EBIT / interest expense

debt_service_coverage:

Source: yfinance

Description: operating cashflow / (interest expense + current portion of long term debt)

cash_coverage:

Source: yfinance

Description: cash and equivalents / (interest expense + current portion of long term debt)

current_ratio:

Source: yfinance

Description: current assets / current liabilities

quick ratio:

Source: yfinance

Description: (cash and equivalents + marketable securities + accounts receivable) / current liabilities

cash ratio:

Source: yfinance

Description: cash and equivalents / current liabilities

Glimpse of Dataset 1

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.3     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
company_credit_risk <- read.csv("company_credit_risk.csv")
glimpse(company_credit_risk)
```

```
Rows: 2,341
Columns: 33
$ X                      <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ~
$ Name                   <chr> "cvs health corp", "cvs health corp", "cvs hea~
$ Sector                 <chr> "Consumer Non-Cyclical", "Consumer Non-Cyclica~
$ CUSIP                  <chr> "126650CZ1", "126650CY4", "126650CX6", "126650~
```

```
$ ISIN                    <chr> "US126650CZ11", "US126650CY46", "US126650CX62"~
$ Price                   <dbl> 84.78, 87.10, 94.64, 85.52, 96.58, 85.03, 95.2~
$ Duration                <dbl> 12.62, 9.77, 3.90, 12.10, 1.69, 3.73, 7.26, 5.~
$ YTM....                 <dbl> 6.27, 6.13, 5.66, 6.36, 5.87, 5.61, 5.91, 5.74~
$ Maturity                <chr> "25-mar-48", "25-mar-38", "25-mar-28", "20-jul~
$ Coupon....              <dbl> 5.05, 4.78, 4.30, 5.13, 3.88, 1.30, 5.25, 3.25~
$ credit_spread           <dbl> 0.93, 0.79, 0.32, 1.02, 0.53, 0.27, 0.57, 0.40~
$ ticker                  <chr> "CVS", "CVS", "CVS", "CVS", "CVS", "CVS", "CVS~
$ marketCapitalization    <dbl> 92335.45, 92335.45, 92335.45, 92335.45, 92335.~
$ shareOutstanding        <dbl> 1284.4, 1284.4, 1284.4, 1284.4, 1284.4, 1284.4~
$ mention                 <dbl> 236, 236, 236, 236, 236, 236, 236, 236, 236, 2~
$ positiveScore           <dbl> 22.18427, 22.18427, 22.18427, 22.18427, 22.184~
$ negativeScore           <dbl> -79.98702, -79.98702, -79.98702, -79.98702, -7~
$ positiveMention         <dbl> 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29, 29~
$ negativeMention         <dbl> 195, 195, 195, 195, 195, 195, 195, 195, 195, 1~
$ score                   <dbl> -59.4749, -59.4749, -59.4749, -59.4749, -59.47~
$ RTG_FITCH               <chr> "", "", "", "", "", "", "", "", "", "", "", ""~
$ RTG_SP                  <chr> "BBB", "BBB", "BBB", "BBB", "BBB", "BBB", "BBB~
$ operating_profit_margin <dbl> 0.8889689, 0.8889689, 0.8889689, 0.8889689, 0.~
$ ebitda_margin           <dbl> 0.04903229, 0.04903229, 0.04903229, 0.04903229~
$ roa                     <dbl> 0.007601841, 0.007601841, 0.007601841, 0.00760~
$ debt_to_assets          <dbl> 0.3284507, 0.3284507, 0.3284507, 0.3284507, 0.~
$ debt_to_equity          <dbl> 1.12939, 1.12939, 1.12939, 1.12939, 1.12939, 1~
$ int_coverage            <dbl> 4.746356, 4.746356, 4.746356, 4.746356, 4.7463~
$ debt_service_coverage   <dbl> 1.913212, 1.913212, 1.913212, 1.913212, 1.9132~
$ cash_coverage           <dbl> 4.471179, 4.471179, 4.471179, 4.471179, 4.4711~
$ current_ratio           <dbl> 0.8554402, 0.8554402, 0.8554402, 0.8554402, 0.~
$ quick_ratio             <dbl> 3.1722e+10, 3.1722e+10, 3.1722e+10, 3.1722e+10~
$ cash_ratio              <dbl> 0.1743176, 0.1743176, 0.1743176, 0.1743176, 0.~
```

Dataset 2: New York Stock Price Exchange

Overview:

The dataset, available on Kaggle, offers a comprehensive collection of historical data pertaining to companies listed on the esteemed New York Stock Exchange (NYSE). This resource proves invaluable for a wide range of financial analyses, encompassing the exploration of stock market trends, performance assessments, and the formulation of effective investment strategies.

The dataset encompasses four distinct CSV files, each providing insight into various aspects of NYSE-listed stock data. Within the "price" and "price-split-adjusted" files, you'll find meticulously recorded daily stock prices for every ticker. The "fundamentals" file meticulously

tracks annual balance sheet fundamentals for each ticker, while the "security" file provides a concise summary of industry categorizations and comprehensive details regarding each ticker's respective company.

Source of Dataset:

Dataset can be found on: NYSE

Prices were fetched from Yahoo Finance, fundamentals are from Nasdaq Financials, extended by some fields from EDGAR SEC databases.

Research Questions:

Research Question 1 (Continuous Outcome Variable):

- **Question**: How is stock price affected by balance sheet fundamentals?
- **Outcome Variable**: `stock price` (numeric and continuous variable; locates in `prices.csv`)
- **Predictor variable**: Balance sheet fundamentals (numeric variable; locates in fundamentals.csv)

Research Question 2 (Ordinal Outcome Variable):

- **Question**: Which sub industry in each sector demonstrates strong correlation between stock price and fundamentals?
- **Outcome Variable**: `Sub industry` (ordinal and categorical variable; locates in `securities.csv`)
- **Predictor variable**: Stock price (numeric and continuous variable; locates in prices.csv)
- **More Predictor variables**: Balance sheet fundamentals (numeric variable; locates in fundamentals.csv)

Dataset 2 Variables Description:

`Price:`

Description:

stock price is recorded daily with the following observations: date, symbol, open, close, low, and high etc. We will only be using the stock price on the last day of each year since this will match the date for balance sheet fundamentals. The mean stock price will be taken from the average of minimum (low) and maximum (high) price.

`Fundamentals (Multiple):`

Description:

Balance sheet fundamentals are recorded for each ticker annually. This includes accounts payable, accounts receivable, cash ratio, changes in inventories, cost of revenue, depreciation, deferred asset charges, deferred liability charges and earnings etc.

`Industry Branch (Multiple):`

Description:

GICS sector and GICS sub industry are recorded for each ticker. These observations identify which industry each ticker belongs to.

Glimpse of Dataset 2

This next section reads the 4 csv files and take a glimpse of each dataset to make sure they are accessible trhough R.

```
security <- read.csv("securities.csv")
price <- read.csv("prices.csv")
price_adjusted <- read.csv("prices-split-adjusted.csv")
fundamentals <- read.csv("fundamentals.csv")
```

```
glimpse(security)
```

```
Rows: 505
Columns: 8
$ Ticker.symbol          <chr> "MMM", "ABT", "ABBV", "ACN", "ATVI", "AYI", "A~
$ Security               <chr> "3M Company", "Abbott Laboratories", "AbbVie",~
$ SEC.filings            <chr> "reports", "reports", "reports", "reports", "r~
$ GICS.Sector            <chr> "Industrials", "Health Care", "Health Care", "~
$ GICS.Sub.Industry      <chr> "Industrial Conglomerates", "Health Care Equip~
```

```
$ Address.of.Headquarters <chr> "St. Paul, Minnesota", "North Chicago, Illinoi~
$ Date.first.added        <chr> "", "1964-03-31", "2012-12-31", "2011-07-06", ~
$ CIK                     <int> 66740, 1800, 1551152, 1467373, 718877, 1144215~
```

```
glimpse(price)
```

```
Rows: 851,264
Columns: 7
$ date   <chr> "2016-01-05 00:00:00", "2016-01-06 00:00:00", "2016-01-07 00:00~
$ symbol <chr> "WLTW", "WLTW", "WLTW", "WLTW", "WLTW", "WLTW", "WLTW", "WLTW",~
$ open   <dbl> 123.43, 125.24, 116.38, 115.48, 117.01, 115.51, 116.46, 113.51,~
$ close  <dbl> 125.84, 119.98, 114.95, 116.62, 114.97, 115.55, 112.85, 114.38,~
$ low    <dbl> 122.31, 119.94, 114.93, 113.50, 114.09, 114.50, 112.59, 110.05,~
$ high   <dbl> 126.25, 125.54, 119.74, 117.44, 117.33, 116.06, 117.07, 115.03,~
$ volume <dbl> 2163600, 2386400, 2489500, 2006300, 1408600, 1098000, 949600, 7~
```

```
glimpse(price_adjusted)
```

```
Rows: 851,264
Columns: 7
$ date   <chr> "2016-01-05", "2016-01-06", "2016-01-07", "2016-01-08", "2016-0~
$ symbol <chr> "WLTW", "WLTW", "WLTW", "WLTW", "WLTW", "WLTW", "WLTW", "WLTW",~
$ open   <dbl> 123.43, 125.24, 116.38, 115.48, 117.01, 115.51, 116.46, 113.51,~
$ close  <dbl> 125.84, 119.98, 114.95, 116.62, 114.97, 115.55, 112.85, 114.38,~
$ low    <dbl> 122.31, 119.94, 114.93, 113.50, 114.09, 114.50, 112.59, 110.05,~
$ high   <dbl> 126.25, 125.54, 119.74, 117.44, 117.33, 116.06, 117.07, 115.03,~
$ volume <dbl> 2163600, 2386400, 2489500, 2006300, 1408600, 1098000, 949600, 7~
```

```
glimpse(fundamentals)
```

```
Rows: 1,781
Columns: 79
$ X                                          <int> 0, 1, 2, 3, 4, 5, ~
$ Ticker.Symbol                              <chr> "AAL", "AAL", "AAL~
$ Period.Ending                              <chr> "2012-12-31", "201~
$ Accounts.Payable                           <dbl> 3068000000, 497500~
$ Accounts.Receivable                        <dbl> -222000000, -93000~
$ Add.l.income.expense.items                 <dbl> -1961000000, -2723~
```

```
$ After.Tax.ROE                                      <dbl> 23, 67, 143, 135, ~
$ Capital.Expenditures                               <dbl> -1888000000, -3114~
$ Capital.Surplus                                    <dbl> 4695000000, 105920~
$ Cash.Ratio                                         <dbl> 53, 75, 60, 51, 23~
$ Cash.and.Cash.Equivalents                          <dbl> 1330000000, 217500~
$ Changes.in.Inventories                             <dbl> 0, 0, 0, 0, -26029~
$ Common.Stocks                                      <dbl> 1.2700e+08, 5.0000~
$ Cost.of.Revenue                                    <dbl> 10499000000, 11019~
$ Current.Ratio                                      <dbl> 78, 104, 88, 73, 1~
$ Deferred.Asset.Charges                             <dbl> 0.000e+00, 0.000e+~
$ Deferred.Liability.Charges                         <dbl> 223000000, 9350000~
$ Depreciation                                       <dbl> 1001000000, 102000~
$ Earnings.Before.Interest.and.Tax                   <dbl> -1813000000, -1324~
$ Earnings.Before.Tax                                <dbl> -2445000000, -2180~
$ Effect.of.Exchange.Rate                            <dbl> 0, 0, 0, 0, 0, 0, ~
$ Equity.Earnings.Loss.Unconsolidated.Subsidiary     <dbl> 0, 0, 0, 0, 0, 0, ~
$ Fixed.Assets                                       <dbl> 13402000000, 19259~
$ Goodwill                                           <dbl> 0, 4086000000, 409~
$ Gross.Margin                                       <dbl> 58, 59, 63, 73, 50~
$ Gross.Profit                                       <dbl> 14356000000, 15724~
$ Income.Tax                                         <dbl> -569000000, -34600~
$ Intangible.Assets                                  <dbl> 869000000, 2311000~
$ Interest.Expense                                   <dbl> 632000000, 8560000~
$ Inventory                                          <dbl> 580000000, 1012000~
$ Investments                                        <dbl> 306000000, -118100~
$ Liabilities                                        <dbl> 473000000, -235000~
$ Long.Term.Debt                                     <dbl> 7116000000, 153530~
$ Long.Term.Investments                              <dbl> 0.00000e+00, 0.000~
$ Minority.Interest                                  <dbl> 0.00e+00, 0.00e+00~
$ Misc..Stocks                                       <dbl> 0.000e+00, 0.000e+~
$ Net.Borrowings                                     <dbl> -1020000000, 22080~
$ Net.Cash.Flow                                      <dbl> 197000000, 6600000~
$ Net.Cash.Flow.Operating                            <dbl> 1285000000, 675000~
$ Net.Cash.Flows.Financing                           <dbl> 483000000, 3799000~
$ Net.Cash.Flows.Investing                           <dbl> -1571000000, -3814~
$ Net.Income                                         <dbl> -1876000000, -1834~
$ Net.Income.Adjustments                             <dbl> 2050000000, 187300~
$ Net.Income.Applicable.to.Common.Shareholders       <dbl> -1876000000, -1834~
$ Net.Income.Cont..Operations                        <dbl> -4084000000, -4489~
$ Net.Receivables                                    <dbl> 1124000000, 156000~
$ Non.Recurring.Items                                <dbl> 386000000, 5590000~
$ Operating.Income                                   <dbl> 148000000, 1399000~
$ Operating.Margin                                   <dbl> 1, 5, 10, 15, 11, ~
```

```
$ Other.Assets                                          <dbl> 2167000000, 229900~
$ Other.Current.Assets                                  <dbl> 626000000, 1465000~
$ Other.Current.Liabilities                             <dbl> 4524000000, 738500~
$ Other.Equity                                          <dbl> -2980000000, -2032~
$ Other.Financing.Activities                            <dbl> 1.5090e+09, 1.7110~
$ Other.Investing.Activities                            <dbl> 11000000, 48100000~
$ Other.Liabilities                                     <dbl> 15147000000, 14915~
$ Other.Operating.Activities                            <dbl> -141000000, -56000~
$ Other.Operating.Items                                 <dbl> 845000000, 8530000~
$ Pre.Tax.Margin                                        <dbl> 10, 8, 8, 11, 10, ~
$ Pre.Tax.ROE                                           <dbl> 31, 80, 159, 82, 5~
$ Profit.Margin                                         <dbl> 8, 7, 7, 19, 6, 6,~
$ Quick.Ratio                                           <dbl> 72, 96, 80, 67, 34~
$ Research.and.Development                              <dbl> 0, 0, 0, 0, 0, 0, ~
$ Retained.Earnings                                     <dbl> -9462000000, -1129~
$ Sale.and.Purchase.of.Stock                            <dbl> 0, 0, -1052000000,~
$ Sales..General.and.Admin.                             <dbl> 12977000000, 12913~
$ Short.Term.Debt...Current.Portion.of.Long.Term.Debt   <dbl> 1419000000, 144600~
$ Short.Term.Investments                                <dbl> 3412000000, 811100~
$ Total.Assets                                          <dbl> 23510000000, 42278~
$ Total.Current.Assets                                  <dbl> 7072000000, 143230~
$ Total.Current.Liabilities                             <dbl> 9011000000, 138060~
$ Total.Equity                                          <dbl> -7987000000, -2731~
$ Total.Liabilities                                     <dbl> 24891000000, 45009~
$ Total.Liabilities...Equity                            <dbl> 16904000000, 42278~
$ Total.Revenue                                         <dbl> 24855000000, 26743~
$ Treasury.Stock                                        <dbl> -367000000, 0, 0, ~
$ For.Year                                              <dbl> 2012, 2013, 2014, ~
$ Earnings.Per.Share                                    <dbl> -5.60, -11.25, 4.0~
$ Estimated.Shares.Outstanding                          <dbl> 335000000, 1630222~
```

Dataset 3: Analyzing an Individual's Credit Default Rate

Overview:

The Home Credit Default Risk dataset originates directly from Home Credit and has been made accessible to the public through a Kaggle competition. This dataset encompasses crucial data from loan applicants, incorporating features significant for credit risk prediction. These features are instrumental in forecasting the likelihood of loan default for each applicant. Obtaining real-world datasets within this domain is typically challenging due to privacy concerns; customer credit data is usually not publicly available. Hence, we have come to rely on this valuable dataset provided by Kaggle.

Within this project, there are eight tables, constituting a total of 346 columns. Our primary focus revolves around utilizing `application_train.csv` and `application_test.csv`, which alone account for 122 columns, for training and testing our model. Other tables, including `bureau.csv`, `bureau_balance.csv`, `POS_CASH_balance.csv`, `credit_card_balance.csv`, `previous_application.csv`, and `installments_payments.csv`, provides specific historical and financial information related to previous credits, loans, and applications of clients, contributing to a comprehensive understanding of their credit behavior and history. So we remain open to extracting variables from other tables if we ascertain their potential utility.

Source of Dataset:

Home Credit Default Risk

Research Questions:

Research Question 1 (Binary Outcome Variable):

- **Question**: Can we predict loan default based on client attributes and financial history?
- **Outcome Variable**: `Target` from **`application_train.csv`**, which is a binary variable indicating loan default (1) or non-default (0) for a given client.

Research Question 2 (Continuous Outcome Variable):

- **Question**: Can we predict total amount of loans a client borrows from Home Credit based on the client's attributes and financial history?
- **Outcome Variable**: Continuous variable representing the total loan amount a person has borrowed, calculated by summing AMT_CREDIT from **`application_train.csv`** and **`previous_application.csv`**.

Dataset 3 Variables Description:

Here are some of the variables we are particularly interested in, and we will keep exploring when we start working on the project:

`AMT_INCOME_TOTAL:`

Description: Total annual income of the applicant.

`AMT_CREDIT:`

Description: Total credit amount applied for.

`AMT_ANNUITY:`

Description: Loan annuity.

NAME_INCOME_TYPE:

Description: Clients' income type.

NAME_CONTRACT_TYPE:

Description: Type of loan: cash loans or revolving loans.

AMT_REQ_CREDIT_BUREAU_YEAR:

Description: Number of enquiries to the Credit Bureau about the client.

FLAG_OWN_CAR and FLAG_OWN_REALTY:

Description: Whether the applicant owns a car or real estate.

REGION_RATING_CLIENT:

Description: Rating of the region where the client lives.

Glimpse of Dataset 3

```
application<- read.csv("application_train.csv")
glimpse(application)
```

```
Rows: 307,511
Columns: 122
$ SK_ID_CURR                  <int> 100002, 100003, 100004, 100006, 100007, 1~
$ TARGET                      <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ NAME_CONTRACT_TYPE          <chr> "Cash loans", "Cash loans", "Revolving lo~
$ CODE_GENDER                 <chr> "M", "F", "M", "F", "M", "M", "F", "M", "~
$ FLAG_OWN_CAR                <chr> "N", "N", "Y", "N", "N", "N", "Y", "Y", "~
$ FLAG_OWN_REALTY             <chr> "Y", "N", "Y", "Y", "Y", "Y", "Y", "Y", "~
$ CNT_CHILDREN                <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,~
$ AMT_INCOME_TOTAL            <dbl> 202500.00, 270000.00, 67500.00, 135000.00~
$ AMT_CREDIT                  <dbl> 406597.5, 1293502.5, 135000.0, 312682.5, ~
$ AMT_ANNUITY                 <dbl> 24700.5, 35698.5, 6750.0, 29686.5, 21865.~
$ AMT_GOODS_PRICE             <dbl> 351000, 1129500, 135000, 297000, 513000, ~
$ NAME_TYPE_SUITE             <chr> "Unaccompanied", "Family", "Unaccompanied~
$ NAME_INCOME_TYPE            <chr> "Working", "State servant", "Working", "W~
$ NAME_EDUCATION_TYPE         <chr> "Secondary / secondary special", "Higher ~
$ NAME_FAMILY_STATUS          <chr> "Single / not married", "Married", "Singl~
$ NAME_HOUSING_TYPE           <chr> "House / apartment", "House / apartment",~
$ REGION_POPULATION_RELATIVE  <dbl> 0.018801, 0.003541, 0.010032, 0.008019, 0~
```

```
$ DAYS_BIRTH                   <int> -9461, -16765, -19046, -19005, -19932, -1~
$ DAYS_EMPLOYED                <int> -637, -1188, -225, -3039, -3038, -1588, -~
$ DAYS_REGISTRATION            <dbl> -3648, -1186, -4260, -9833, -4311, -4970,~
$ DAYS_ID_PUBLISH              <int> -2120, -291, -2531, -2437, -3458, -477, -~
$ OWN_CAR_AGE                  <dbl> NA, NA, 26, NA, NA, NA, 17, 8, NA, NA, NA~
$ FLAG_MOBIL                   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ FLAG_EMP_PHONE               <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1,~
$ FLAG_WORK_PHONE              <int> 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0,~
$ FLAG_CONT_MOBILE             <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
$ FLAG_PHONE                   <int> 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0,~
$ FLAG_EMAIL                   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ OCCUPATION_TYPE              <chr> "Laborers", "Core staff", "Laborers", "La~
$ CNT_FAM_MEMBERS              <dbl> 1, 2, 1, 2, 1, 2, 3, 2, 2, 1, 3, 2, 2, 3,~
$ REGION_RATING_CLIENT         <int> 2, 1, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2,~
$ REGION_RATING_CLIENT_W_CITY  <int> 2, 1, 2, 2, 2, 2, 2, 3, 2, 2, 2, 2, 2, 2,~
$ WEEKDAY_APPR_PROCESS_START   <chr> "WEDNESDAY", "MONDAY", "MONDAY", "WEDNESD~
$ HOUR_APPR_PROCESS_START      <int> 10, 11, 9, 17, 11, 16, 16, 16, 14, 8, 15,~
$ REG_REGION_NOT_LIVE_REGION   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ REG_REGION_NOT_WORK_REGION   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ LIVE_REGION_NOT_WORK_REGION  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ REG_CITY_NOT_LIVE_CITY       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ REG_CITY_NOT_WORK_CITY       <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
$ LIVE_CITY_NOT_WORK_CITY      <int> 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0,~
$ ORGANIZATION_TYPE            <chr> "Business Entity Type 3", "School", "Gove~
$ EXT_SOURCE_1                 <dbl> 0.08303697, 0.31126731, NA, NA, NA, NA, 0~
$ EXT_SOURCE_2                 <dbl> 0.2629486, 0.6222458, 0.5559121, 0.650441~
$ EXT_SOURCE_3                 <dbl> 0.13937578, NA, 0.72956669, NA, NA, 0.621~
$ APARTMENTS_AVG               <dbl> 0.0247, 0.0959, NA, NA, NA, NA, NA, NA, N~
$ BASEMENTAREA_AVG             <dbl> 0.0369, 0.0529, NA, NA, NA, NA, NA, NA, N~
$ YEARS_BEGINEXPLUATATION_AVG  <dbl> 0.9722, 0.9851, NA, NA, NA, NA, NA, NA, N~
$ YEARS_BUILD_AVG              <dbl> 0.6192, 0.7960, NA, NA, NA, NA, NA, NA, N~
$ COMMONAREA_AVG               <dbl> 0.0143, 0.0605, NA, NA, NA, NA, NA, NA, N~
$ ELEVATORS_AVG                <dbl> 0.00, 0.08, NA, NA, NA, NA, NA, NA, NA, N~
$ ENTRANCES_AVG                <dbl> 0.0690, 0.0345, NA, NA, NA, NA, NA, NA, N~
$ FLOORSMAX_AVG                <dbl> 0.0833, 0.2917, NA, NA, NA, NA, NA, NA, N~
$ FLOORSMIN_AVG                <dbl> 0.1250, 0.3333, NA, NA, NA, NA, NA, NA, N~
$ LANDAREA_AVG                 <dbl> 0.0369, 0.0130, NA, NA, NA, NA, NA, NA, N~
$ LIVINGAPARTMENTS_AVG         <dbl> 0.0202, 0.0773, NA, NA, NA, NA, NA, NA, N~
$ LIVINGAREA_AVG               <dbl> 0.0190, 0.0549, NA, NA, NA, NA, NA, NA, N~
$ NONLIVINGAPARTMENTS_AVG      <dbl> 0.0000, 0.0039, NA, NA, NA, NA, NA, NA, N~
$ NONLIVINGAREA_AVG            <dbl> 0.0000, 0.0098, NA, NA, NA, NA, NA, NA, N~
$ APARTMENTS_MODE              <dbl> 0.0252, 0.0924, NA, NA, NA, NA, NA, NA, N~
$ BASEMENTAREA_MODE            <dbl> 0.0383, 0.0538, NA, NA, NA, NA, NA, NA, N~
```

```
$ YEARS_BEGINEXPLUATATION_MODE <dbl> 0.9722, 0.9851, NA, NA, NA, NA, NA, NA, N~
$ YEARS_BUILD_MODE             <dbl> 0.6341, 0.8040, NA, NA, NA, NA, NA, NA, N~
$ COMMONAREA_MODE              <dbl> 0.0144, 0.0497, NA, NA, NA, NA, NA, NA, N~
$ ELEVATORS_MODE               <dbl> 0.0000, 0.0806, NA, NA, NA, NA, NA, NA, N~
$ ENTRANCES_MODE               <dbl> 0.0690, 0.0345, NA, NA, NA, NA, NA, NA, N~
$ FLOORSMAX_MODE               <dbl> 0.0833, 0.2917, NA, NA, NA, NA, NA, NA, N~
$ FLOORSMIN_MODE               <dbl> 0.1250, 0.3333, NA, NA, NA, NA, NA, NA, N~
$ LANDAREA_MODE                <dbl> 0.0377, 0.0128, NA, NA, NA, NA, NA, NA, N~
$ LIVINGAPARTMENTS_MODE        <dbl> 0.0220, 0.0790, NA, NA, NA, NA, NA, NA, N~
$ LIVINGAREA_MODE              <dbl> 0.0198, 0.0554, NA, NA, NA, NA, NA, NA, N~
$ NONLIVINGAPARTMENTS_MODE     <dbl> 0.0000, 0.0000, NA, NA, NA, NA, NA, NA, N~
$ NONLIVINGAREA_MODE           <dbl> 0.0000, 0.0000, NA, NA, NA, NA, NA, NA, N~
$ APARTMENTS_MEDI              <dbl> 0.0250, 0.0968, NA, NA, NA, NA, NA, NA, N~
$ BASEMENTAREA_MEDI            <dbl> 0.0369, 0.0529, NA, NA, NA, NA, NA, NA, N~
$ YEARS_BEGINEXPLUATATION_MEDI <dbl> 0.9722, 0.9851, NA, NA, NA, NA, NA, NA, N~
$ YEARS_BUILD_MEDI             <dbl> 0.6243, 0.7987, NA, NA, NA, NA, NA, NA, N~
$ COMMONAREA_MEDI              <dbl> 0.0144, 0.0608, NA, NA, NA, NA, NA, NA, N~
$ ELEVATORS_MEDI               <dbl> 0.00, 0.08, NA, NA, NA, NA, NA, NA, NA, N~
$ ENTRANCES_MEDI               <dbl> 0.0690, 0.0345, NA, NA, NA, NA, NA, NA, N~
$ FLOORSMAX_MEDI               <dbl> 0.0833, 0.2917, NA, NA, NA, NA, NA, NA, N~
$ FLOORSMIN_MEDI               <dbl> 0.1250, 0.3333, NA, NA, NA, NA, NA, NA, N~
$ LANDAREA_MEDI                <dbl> 0.0375, 0.0132, NA, NA, NA, NA, NA, NA, N~
$ LIVINGAPARTMENTS_MEDI        <dbl> 0.0205, 0.0787, NA, NA, NA, NA, NA, NA, N~
$ LIVINGAREA_MEDI              <dbl> 0.0193, 0.0558, NA, NA, NA, NA, NA, NA, N~
$ NONLIVINGAPARTMENTS_MEDI     <dbl> 0.0000, 0.0039, NA, NA, NA, NA, NA, NA, N~
$ NONLIVINGAREA_MEDI           <dbl> 0.0000, 0.0100, NA, NA, NA, NA, NA, NA, N~
$ FONDKAPREMONT_MODE           <chr> "reg oper account", "reg oper account", "~
$ HOUSETYPE_MODE               <chr> "block of flats", "block of flats", "", "~
$ TOTALAREA_MODE               <dbl> 0.0149, 0.0714, NA, NA, NA, NA, NA, NA, N~
$ WALLSMATERIAL_MODE           <chr> "Stone, brick", "Block", "", "", "", "", ~
$ EMERGENCYSTATE_MODE          <chr> "No", "No", "", "", "", "", "", "", "", "~
$ OBS_30_CNT_SOCIAL_CIRCLE     <dbl> 2, 1, 0, 2, 0, 0, 1, 2, 1, 2, 0, 0, 0, 0,~
$ DEF_30_CNT_SOCIAL_CIRCLE     <dbl> 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ OBS_60_CNT_SOCIAL_CIRCLE     <dbl> 2, 1, 0, 2, 0, 0, 1, 2, 1, 2, 0, 0, 0, 0,~
$ DEF_60_CNT_SOCIAL_CIRCLE     <dbl> 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ DAYS_LAST_PHONE_CHANGE       <dbl> -1134, -828, -815, -617, -1106, -2536, -1~
$ FLAG_DOCUMENT_2              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_3              <int> 1, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1,~
$ FLAG_DOCUMENT_4              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_5              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_6              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,~
$ FLAG_DOCUMENT_7              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_8              <int> 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
```

```
$ FLAG_DOCUMENT_9              <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_10             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_11             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_12             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_13             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_14             <int> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_15             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_16             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_17             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_18             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_19             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_20             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ FLAG_DOCUMENT_21             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
$ AMT_REQ_CREDIT_BUREAU_HOUR   <dbl> 0, 0, 0, NA, 0, 0, 0, 0, 0, NA, 0, 0, 0, ~
$ AMT_REQ_CREDIT_BUREAU_DAY    <dbl> 0, 0, 0, NA, 0, 0, 0, 0, 0, NA, 0, 0, 0, ~
$ AMT_REQ_CREDIT_BUREAU_WEEK   <dbl> 0, 0, 0, NA, 0, 0, 0, 0, 0, NA, 0, 0, 0, ~
$ AMT_REQ_CREDIT_BUREAU_MON    <dbl> 0, 0, 0, NA, 0, 0, 1, 0, 0, NA, 1, 0, 1, ~
$ AMT_REQ_CREDIT_BUREAU_QRT    <dbl> 0, 0, 0, NA, 0, 1, 1, 0, 0, NA, 0, 0, 0, ~
$ AMT_REQ_CREDIT_BUREAU_YEAR   <dbl> 1, 0, 0, NA, 0, 1, 2, 0, 1, NA, 0, 2, 0, ~
```