

Отчет по Заданию №3

По предоставленным историческим данным в виде таблицы *loanform_features.csv*

1. Определить какие факторы влияют на вероятность невозврата кредита и отранжировать их по прогностической силе.
2. Определить какие факторы влияют на вероятность одобрения кредита и отранжировать их по прогностической силе.

1. Для анализа данных принято решение использовать пакет *pandas*, в виду его возможностей в области проведения расчетов над табличными данными и визуализации, а так же потому что *pandas* является высокоуровневым пакетом и позволяет минимизировать время на разработку кода.

Начинаем анализ с загрузки данных и визуального анализа *DataFrame*. (код модуля в репозитории, в папке *app*, под названием *predict_power_factors.py* : https://github.com/mark-rtb/data_mining.git)

Index	ORDERID	APPROVED	ISSUED	BAD	STRAFDAYSQUAN	ORDERSTATUS	FAMILYQUANT	FAMILYSTATUS	CLIENTAGE	MONTHINCOME	FIRSTLOAN	issued_sum	IS_ONLINE
0	2528520	отказано	не выдан	nan	0	3	0	6	42	51	10000	0	1
1	2528521	отказано	не выдан	nan	0	3	0	6	21	6	100000	0	1
2	2528522	отказано	не выдан	nan	0	3	0	1	32	31	2000	0	1
3	2528523	отказано	не выдан	nan	0	3	0	5	33	80	4000	0	1
4	2528524	отказано	не выдан	nan	0	3	0	1	38	25	10000	0	1

Рис.1. Первые пять строк *DataFrame*.

Как видно из рисунка 1, данные для анализа имеют как категориальные, так и количественные признаки. Так же имеются пропуски в данных(*nan*).

Ключевым признаком для анализа возврата кредита является столбец "*BAD*".

Удалим строки в *DataFrame* в которых в столбце "*BAD*", значение *nan*.

Удалим признаки '*ORDERID*', '*APPROVED*', '*ISSUED*' так как очевидно, что они не могут повлиять на возврат кредита, удалим столбец '*STRAFDAYSQUAN*' так как на момент принятия решения у нас не может быть этих данных.

Заменим в столбце "*BAD*", значения «вернул», «не вернул» на 1 и 0 соответственно.

Посчитаем корреляцию между парами значений и построим их зависимости.

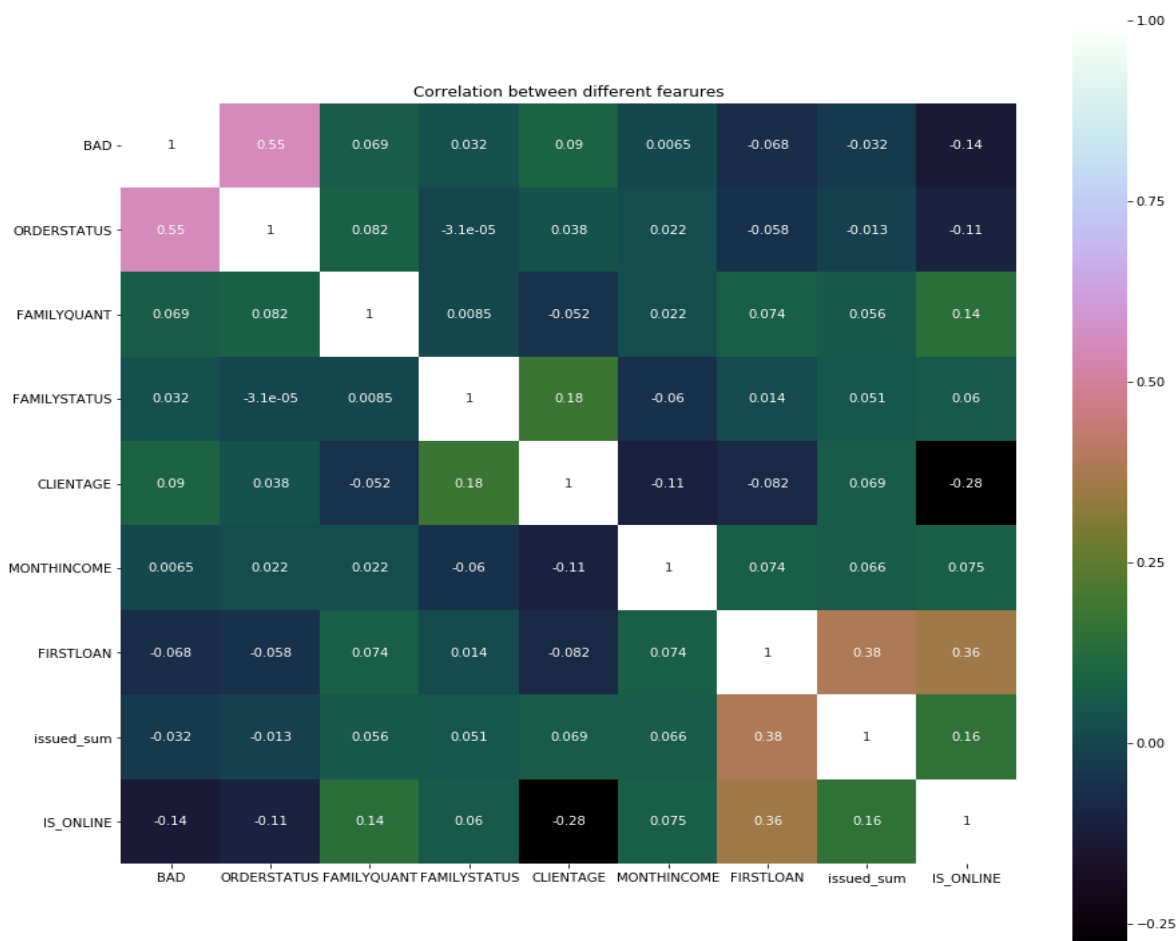


Рис.2. Корреляционные зависимости пар факторов.

Визуальный анализ рисунка 2 показывает очевидную зависимость возврата кредита от таких факторов как *SHTRAFDAYSQUANT*, *ORDERSTATUS*, *IS_ONLINE* и т.д., однако эксперты отмечают, что не всегда высокая корреляция является показателем прогностической силы фактора. В научном сообществе рекомендуется для оценки использовать мера прогностической силы переменной - *information value (IV)*.

Рассчитаем *information value (IV)*, как видно из таблицы некоторые параметры поменялись местами по прогностической силе, в сравнении с корреляционными коэффициентами. Так же немаловажен тот факт, что *information value (IV)*, позволяет численно оценить влияние факторов.

VAR_NAME	IV
FIRSTLOAN	0.172628
IS_ONLINE	0.103522
CLIENTAGE	0.046501
ORDERSTATUS	0.015547
issued_sum	0.010820
FAMILYQUANT	0.007275
FAMILYSTATUS	0.003647
MONTHINCOME	0.000428

Табл.1. Оценки информативности переменной методом *information value (IV)*.

2. Выполнив все те же шаги для фактора одобрения кредита, получим корреляционную зависимость и оценку меры прогностической силы методом *information value (IV)*:

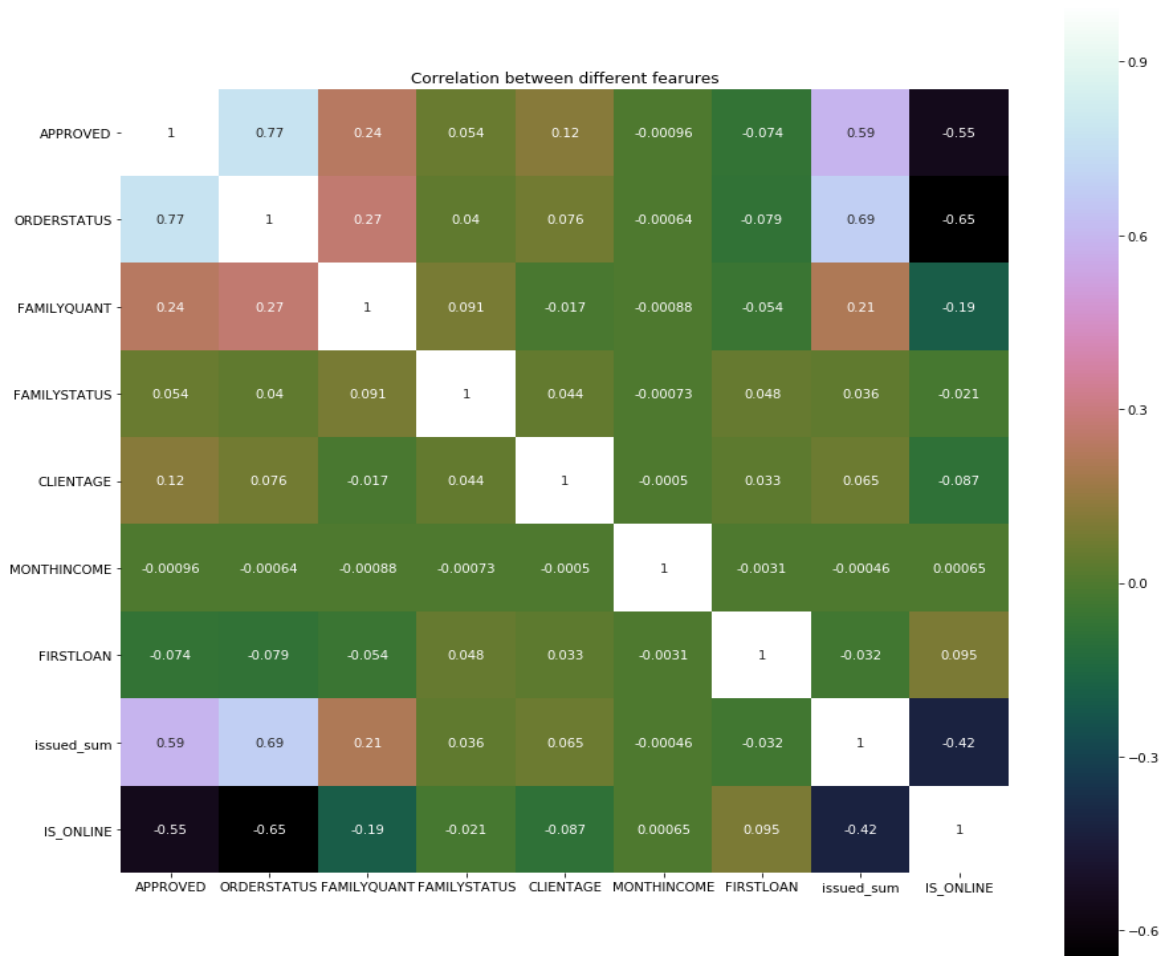


Рис.3. Корреляционные зависимости пар факторов.

VAR_NAME	IV
IS_ONLINE	1.957993
ORDERSTATUS	0.740756
issued_sum	0.740756
FAMILYQUANT	0.673922
MONTHINCOME	0.493780
FAMILYSTATUS	0.345741
CLIENTAGE	0.300749
FIRSTLOAN	0.171032

Табл.2. Оценки информативности переменной методом *information value (IV)*.

В процессе анализа модулем из репозитория создаются объекты *.pkl* в дирректории с данными, которые в дальнейшем можно использовать для построения моделей машинного обучения.

Код модуля написан на языке *Python 3.6*, для корректной работы модуля необходимы библиотеки: *Pandas, seaborn, matplotlib, os, numpy, scipy, re*