ЦМФ 2021. Bank Scoring Case, level 2
Predicting probability of default

Работу выполнил:

Медведев Виктор

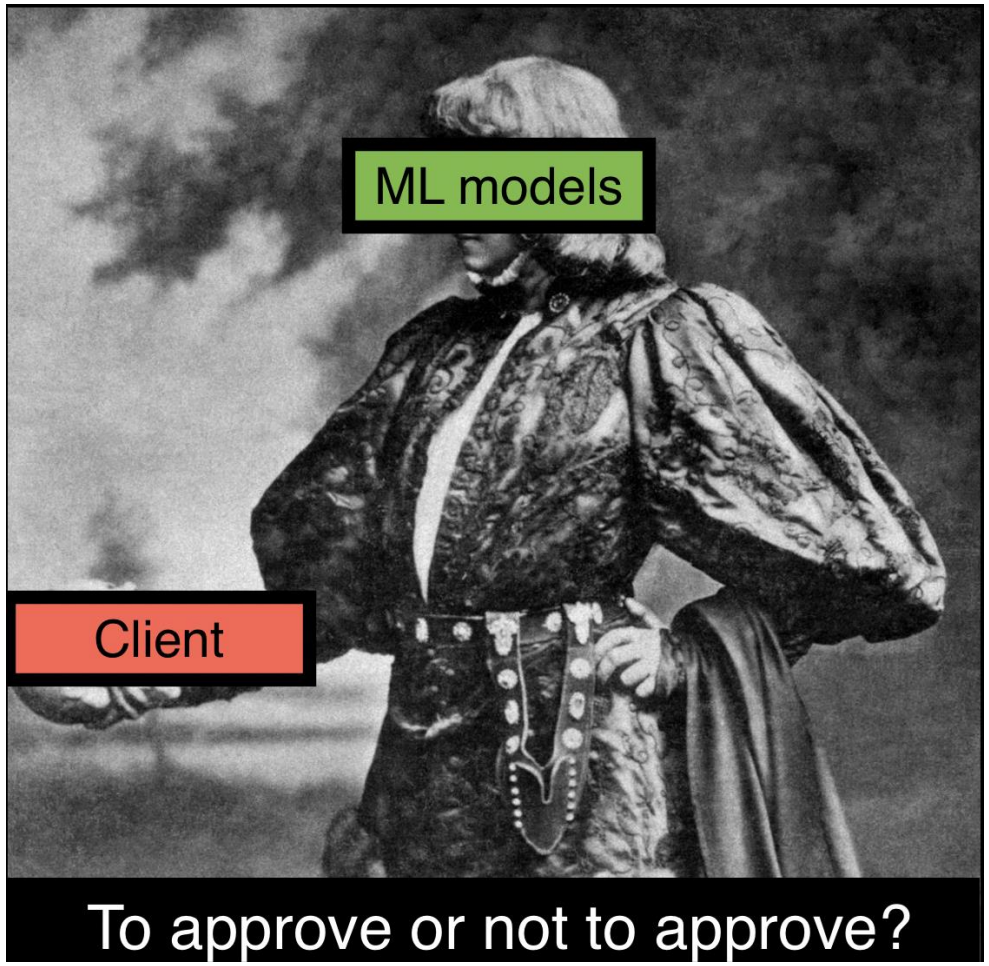
Постановка задачи

Бизнес постановка:

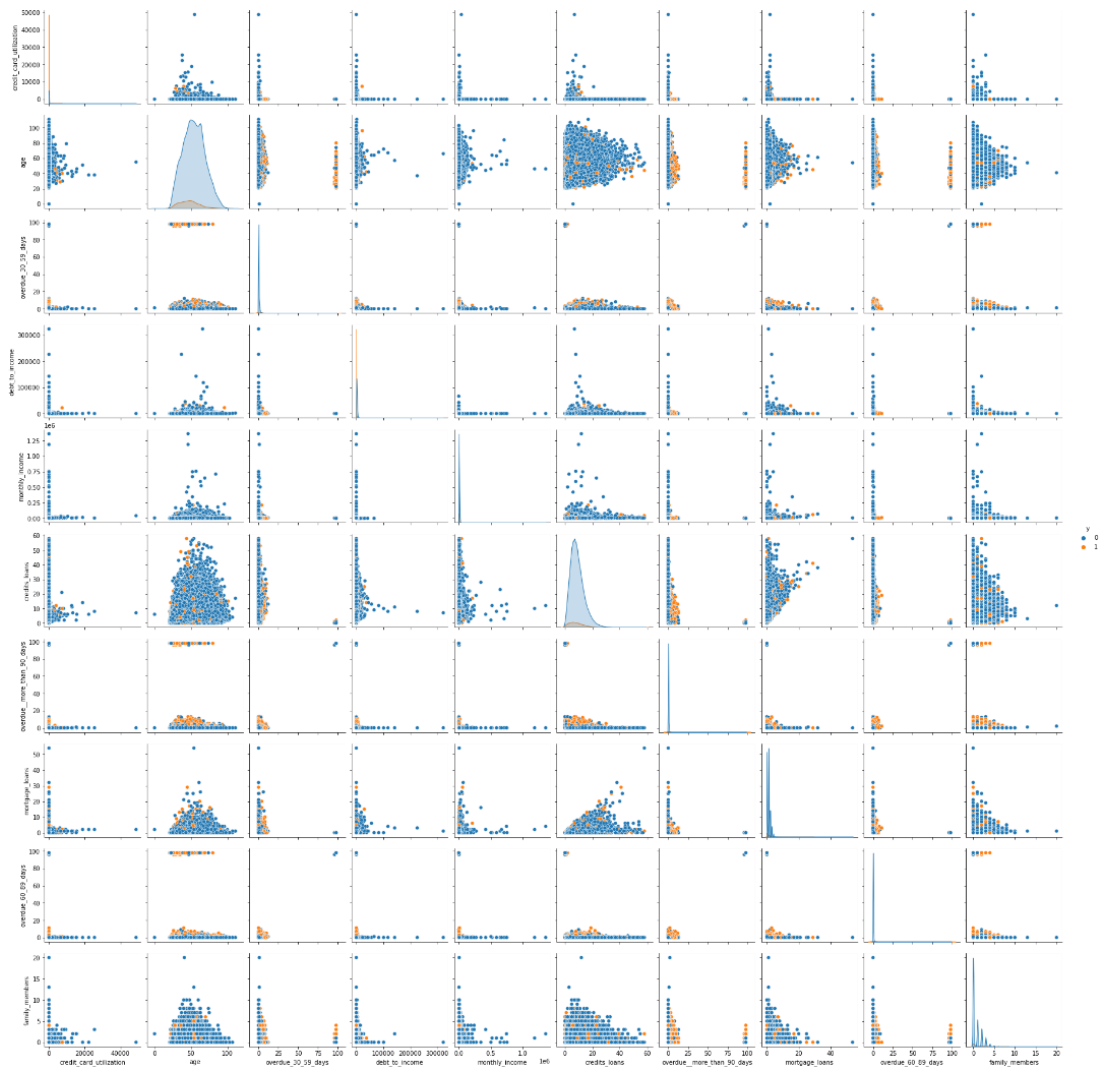
- выборка с данными клиентов
- необходимо предсказать, произойдет ли просрочка кредита у отдельно взятого клиента

Постановка машинного обучения:

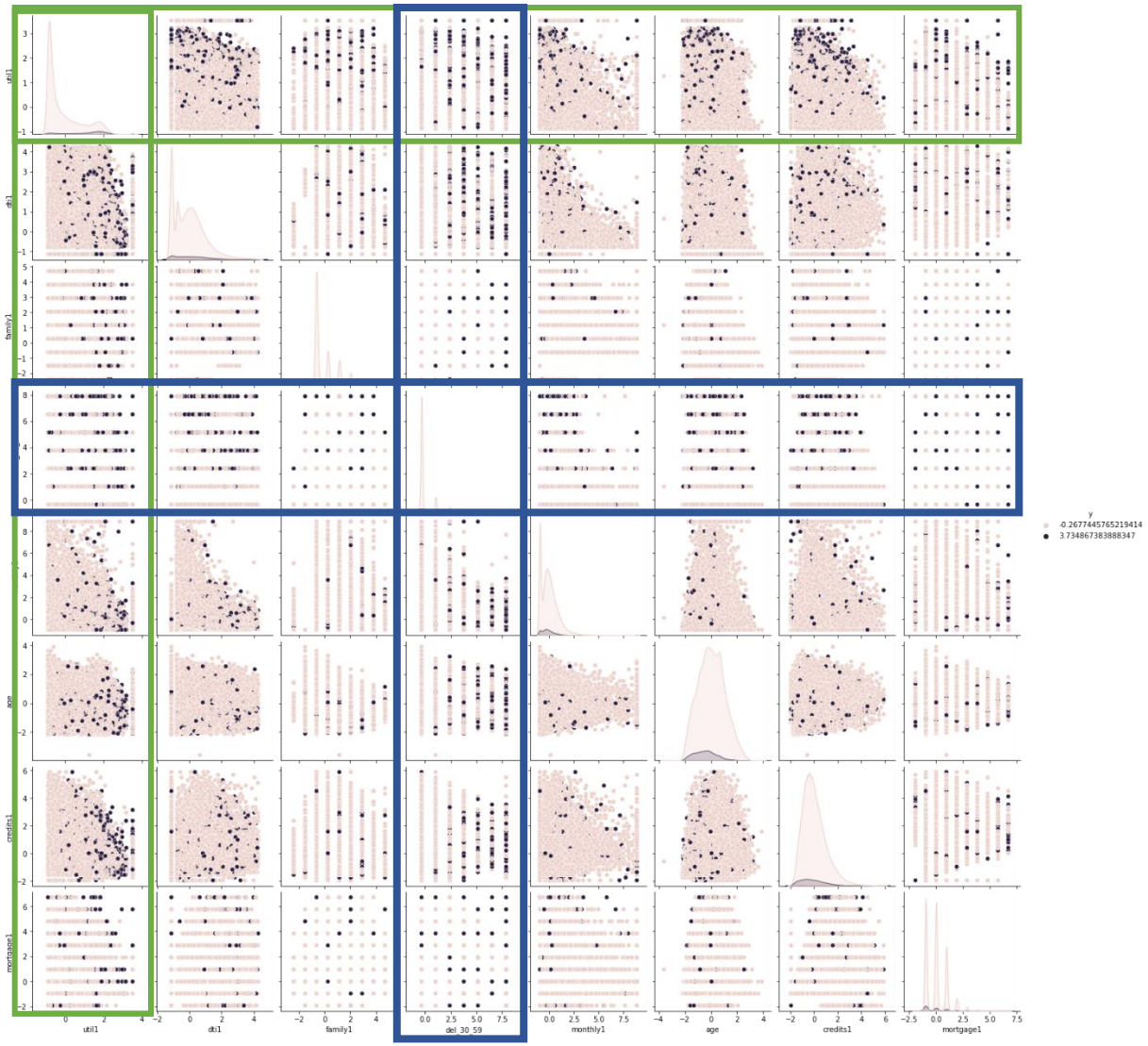
- данные в X_train, X_test
- известный таргет в y_train, предсказать y_test
- анализ данных и их предобработка
- тип задачи - задача классификации с учителем
- тестирование моделей и выбор лучшей



Визуализация исходных и обработанных данных



Исходные данные: множество выбросов,
ни один из признаков визуально не разделяет классы



Обработанные данные: нет выбросов,
заметна разделяющая сила утилизации и дней просрочки

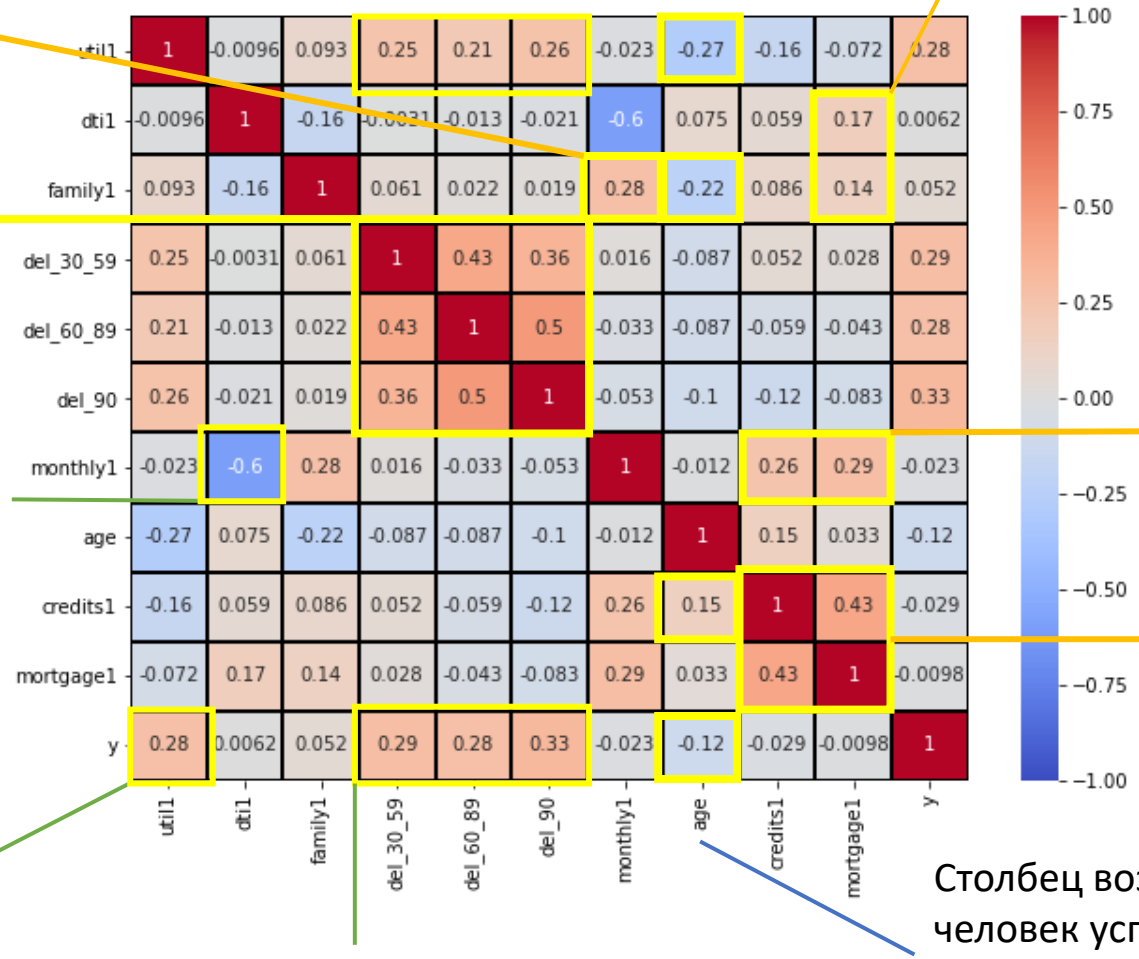
Матрица корреляции

Чем больше доход, тем больше число членов семьи ☺

Просрочка коррелирует не только с таргетом, но и с утилизацией

Чем больше долговая нагрузка и число членов семьи, тем нужнее человеку ипотека

Взаимная корреляция разных видов просрочки



Сильная отрицательная корреляция DTI с доходом

Высокий доход позволяет взять больше ипотек и кредитов

Чем больше у человека ипотек, тем больше кредитов

Высокая корреляция утилизации с таргетом

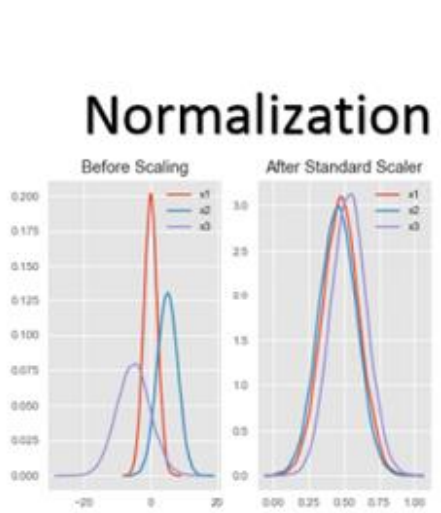
Высокая корреляция просрочки с таргетом

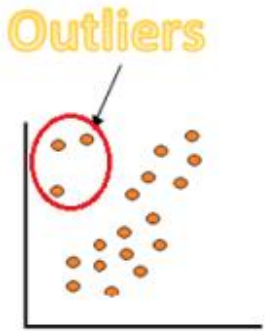
Столбец возраст. Чем старше, тем больше человек успевает взять кредитов, меньше дефолтит, у него ниже утилизация, а также число членов семьи ☹

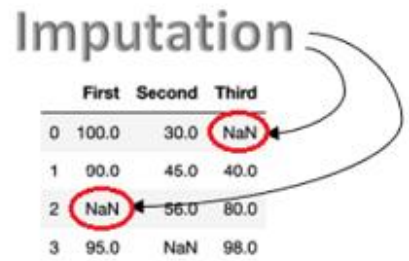
Предобработка данных

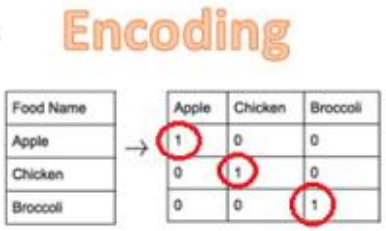
После визуализации всех зависимостей в данных:

- Необходимо посмотреть важные зависимости между фичами и target значениями
- Определить статистическую значимость значений фичей
- Найти явные ошибки в данных
- Заполнить пропущенные значения в данных

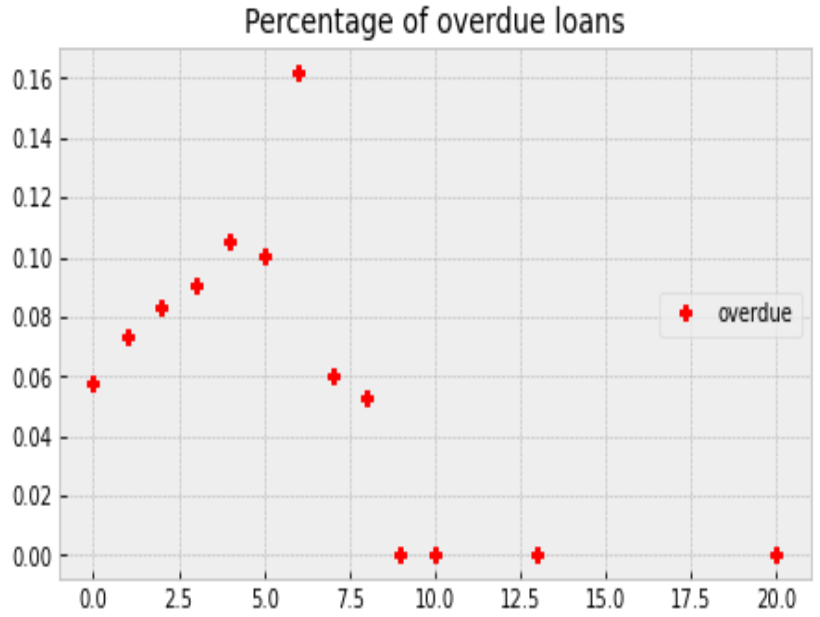


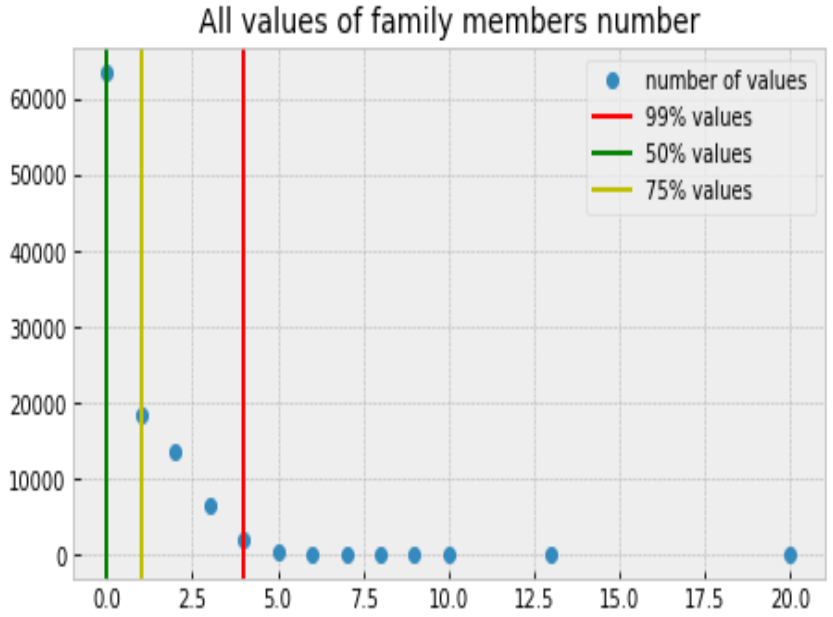






Предобработка данных





Family members содержит пропущенные значения, для заполнения пропущенных значений использовался метод **most frequent.**

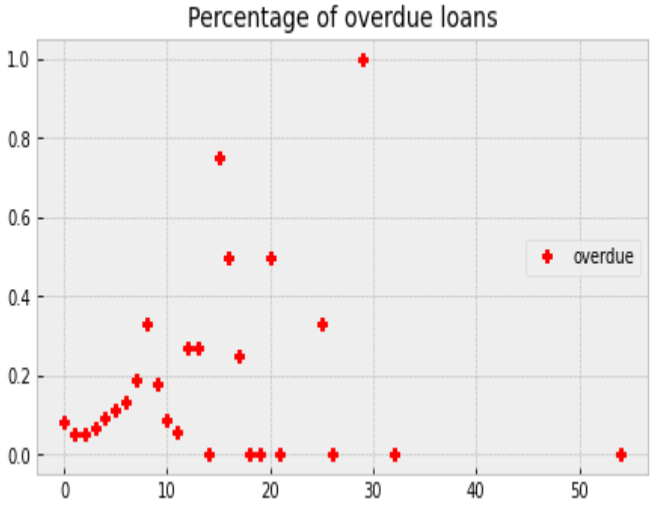
На диаграмме справа изображено количество повторяющихся значений числа членов семьи, а также отмечены квантили (50%, 75% и 99%).

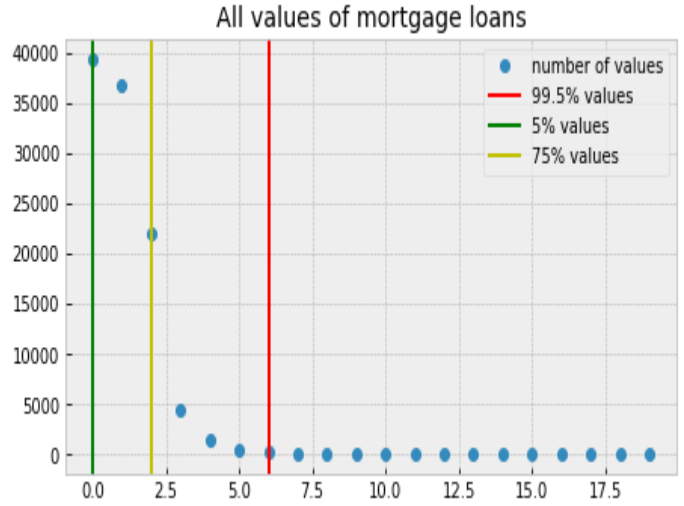
Как видно, семьи с числом членов больше 4 можно удалить из данных, как статистически не значимые и таким образом мы исключим ложные зависимости в данных, как на левом графике (в его правой части).

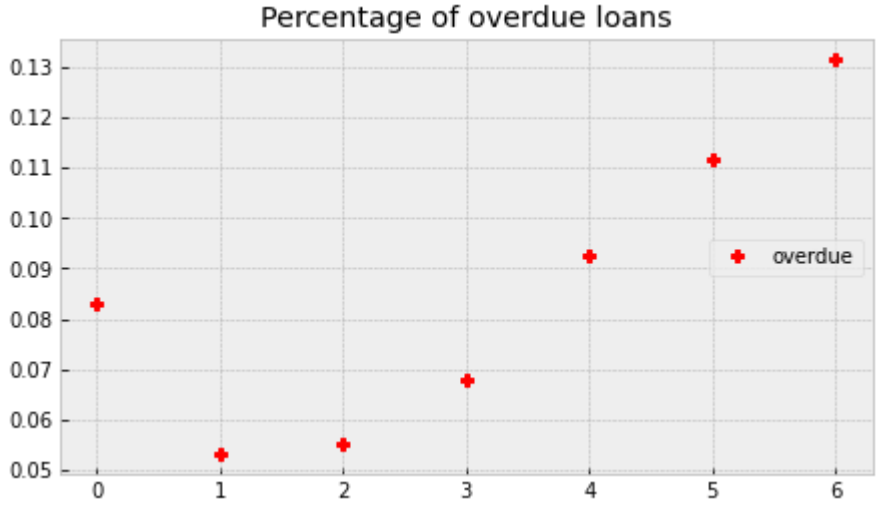
Предобработка данных

Mortgage loans не содержит пропуски, поэтому нужно проанализировать только зависимости и их статистическую значимость. На графике слева (рис. 2) изображены значения вероятности просрочки по кредиту в зависимости от текущего количества ипотечных займов. После значения 9 наблюдается довольная странная зависимость, но после анализа выборки по правой диаграмме, становится понятно, что значения больше 6, статистически не значимы, а потому их необходимо выбросить.

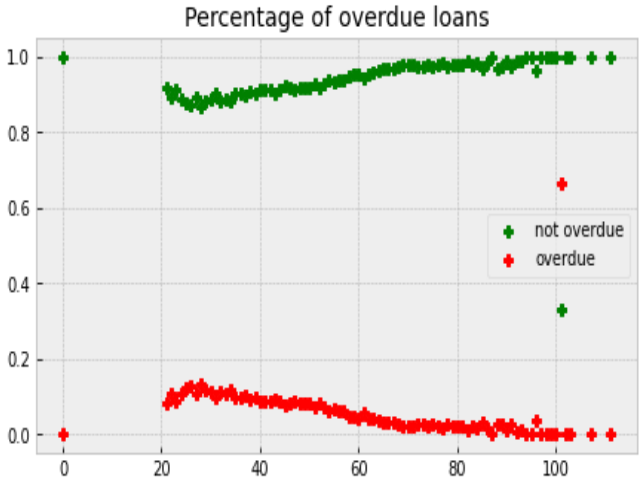
После того, как они были удалены, остается только следующая зависимость в данных

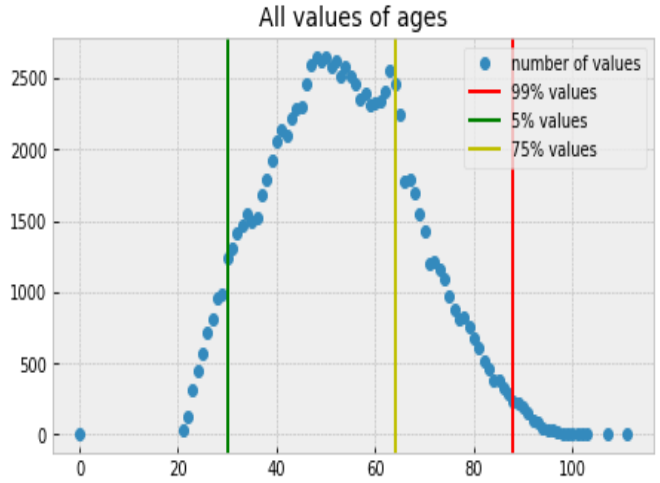




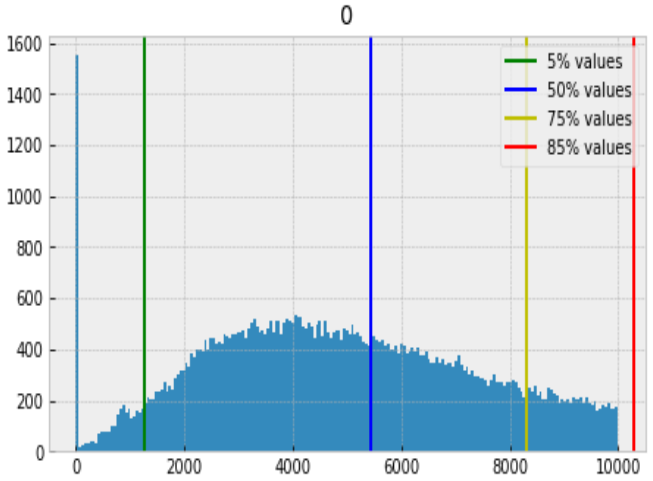


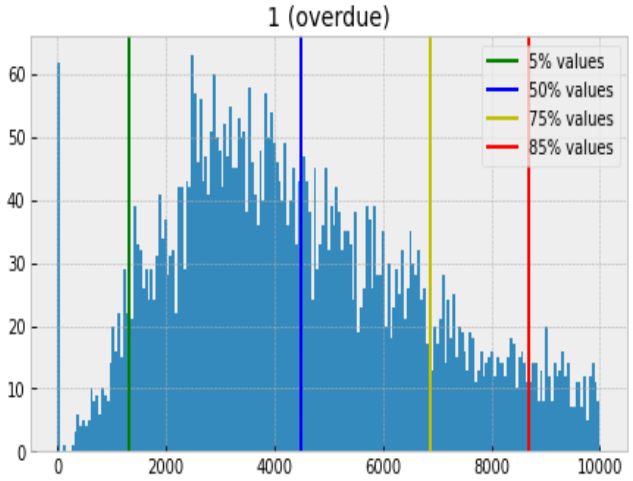
Предобработка данных





Видно, что на этой диаграмме содержится значение возраста 0, которое мы тоже должны выбросить.

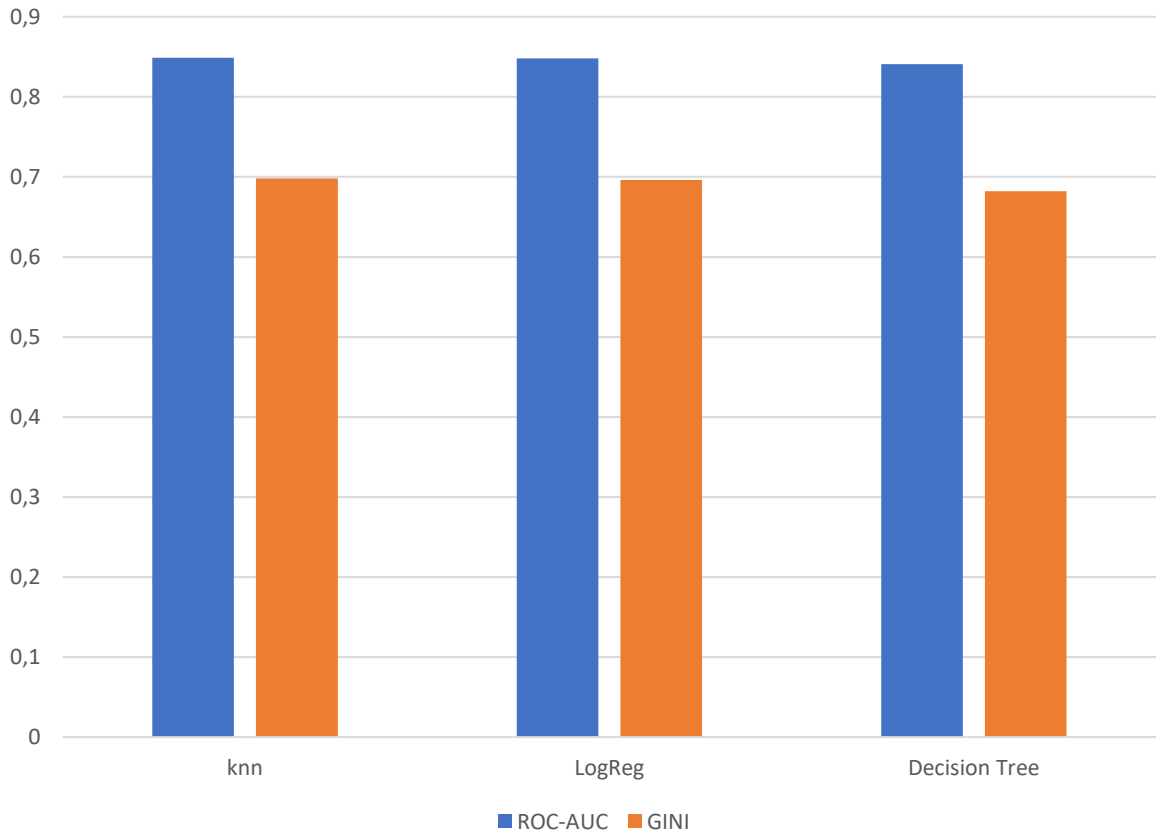




На левой гистограмме по **monthly income** те, кто вовремя возвращал кредит, а на правой те, кто допускал просрочки.

Как видно из гистограмм, левая медиана на значении чуть больше 5000, а правая медиана на значении чуть меньше 5000, поэтому пропущенные значения заполнялись посредством среднего значения этих двух медиан.

Простые модели



	ROC-AUC	GINI
knn	0,849	0,698
LogReg	0,848	0,696
Decision Tree	0,841	0,682

- **Наилучшие модели:** LogReg и knn

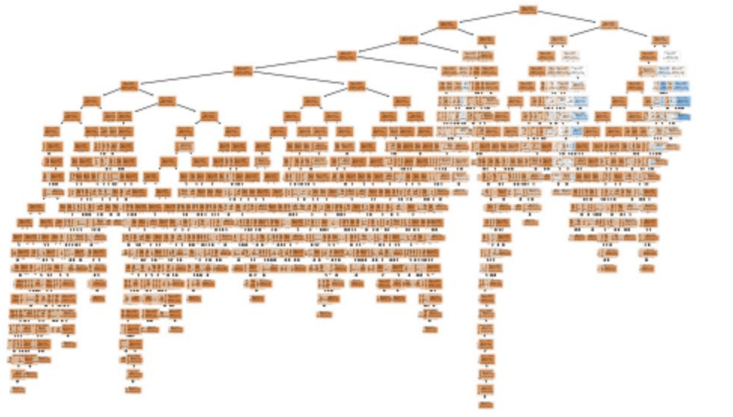
- В LogReg нужно избежать мультиколлинеарности, зато она стабильна, интерпретируема и преобразуема к классическим скоринговым моделям
- KNN устойчив к выбросам, достаточно прост и интерпретируем, но трудоемок и затратен в вычислительном плане
- Decision Tree обладает высокой объясняющей способностью, требует небольшой предобработки, но неустойчив к изменениям

Random Forest

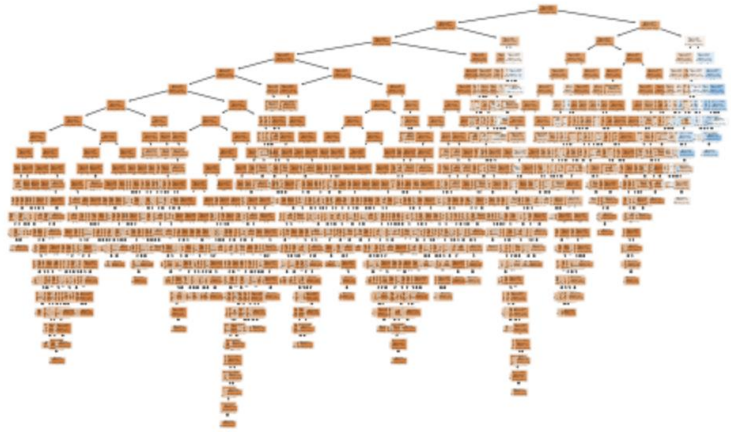
Лучшие параметры:

max_depth=90, max_features=2, min_samples_leaf=40,
min_samples_split=10, n_estimators=200

Результат: 0.85861



first tree

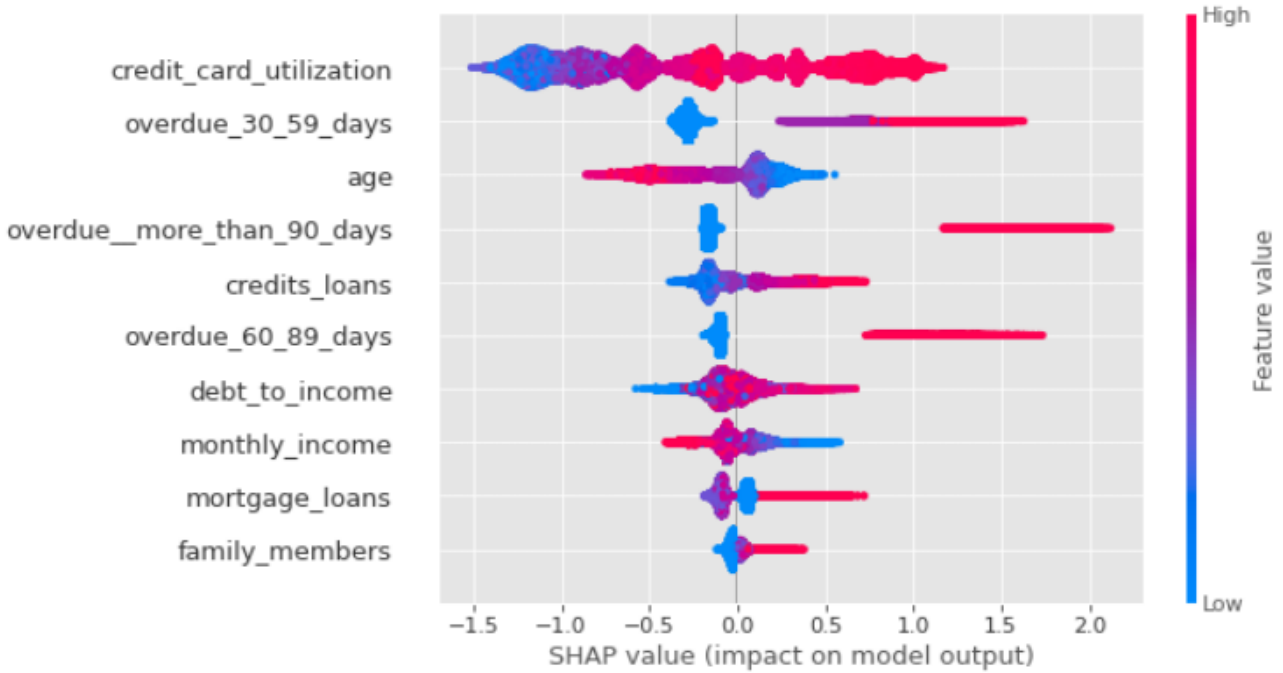


second tree

third tree

и еще 197 таких решающих деревьев...

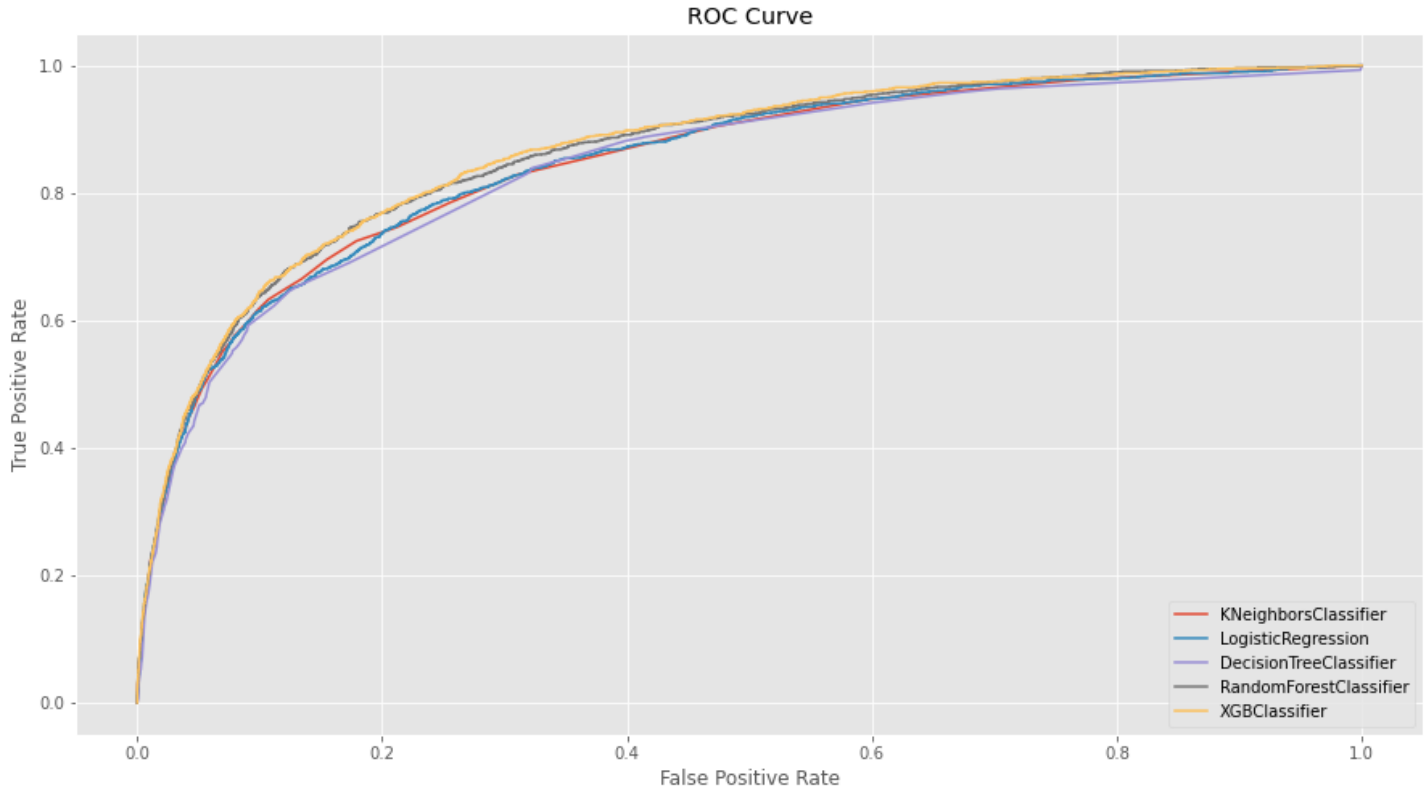
XGBoost



Model Report
Accuracy : 0.9372
Precision : 0.5984
Recall : 0.1909
F1 score : 0.2894
AUC Score (Train): 0.8708
AUC Score (Test): 0.8656
AUC Score (Kaggle): 0.8652

Лучшие параметры модели:
learning_rate=0.1, n_estimators=150, max_depth=3, min_child_weight=4, gamma=0.6, subsample=0.6, colsample_bytree=0.8, reg_alpha=1, objective= 'binary:logistic', nthread=4, scale_pos_weight=1

Анализ метрик качества



Logistic Regression:
precision - 0.8355;
recall - 0.9743;
f1 - 0.8996;
accuracy - 0.8258;
AUC-ROC - 0.8492;
AUC-ROC (Kaggle) - 0.8486

kNN:
precision - 0.7881;
recall - 0.9776;
f1 - 0.8727;
accuracy - 0.7853;
AUC-ROC - 0.8484;
AUC-ROC (Kaggle) - 0.8421

Decision Tree Classifier:
precision - 0.8698;
recall - 0.9724;
f1 - 0.9182;
accuracy - 0.8554;
AUC-ROC - 0.8416;
AUC-ROC (Kaggle) - 0.8344

Random Forest Classifier:
precision - 0.7825;
recall - 0.9808;
f1 - 0.8705;
accuracy - 0.7826;
AUC-ROC - 0.8622;
AUC-ROC (Kaggle) - 0.861