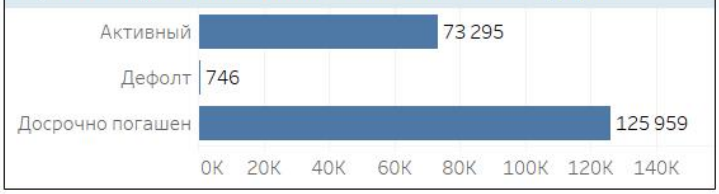
Первичный анализ (EDA) ипотечных дефолтов.

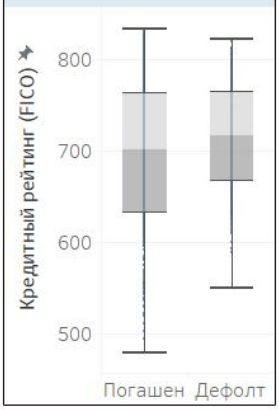
Этот дэшборд является частью проекта по анализу исторических данных по ипотечным займам от Freddie Mac (2017–2020). Цель — выявить ключевые факторы, связанные с дефолтами по ипотеке, и подготовить данные к построению прогностической модели. На текущем этапе представлен исследовательский анализ (EDA), который выявил:

- сильный дисбаланс классов.
- заметные различия в доле дефолтов между типами недвижимости, регионами и обслуживающими организациями,
- неожиданный результат по кредитному рейтингу (у дефолтов он не ниже..

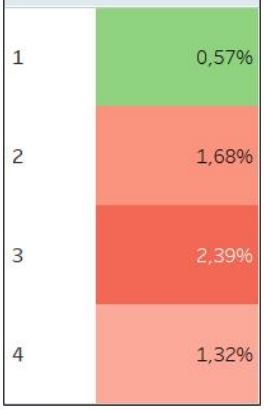
Выраженный дисбаланс классов: 170 погашенных на 1 дефолт



Кредитный рейтинг у дефолтов — сопоставим или даже чуть выше

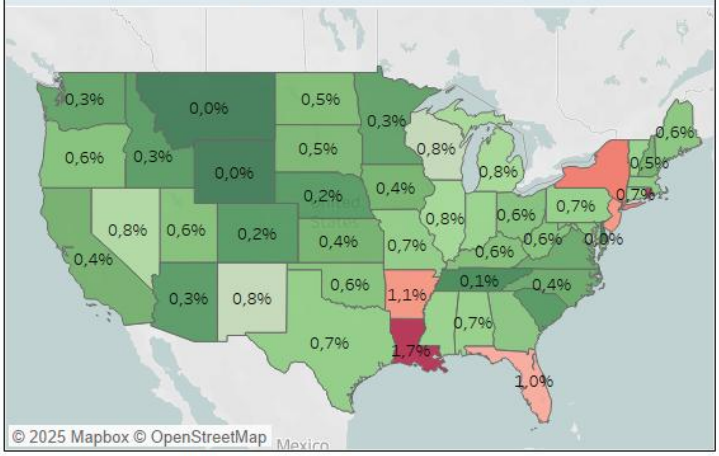


Доля дефолтов по числу квартир, покрытых одной ипотекой

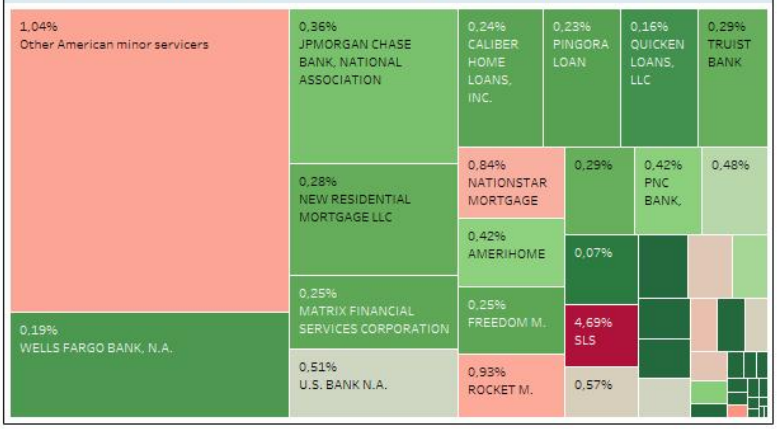


Доля дефолтов по типу недвижимости

География дефолтов: уровень по штатам США



Разный объём и разный риск: обслуживающие банки отличаются и по числу выданных ипотек (размер прямоугольника), и по доле дефолтов (цвет)

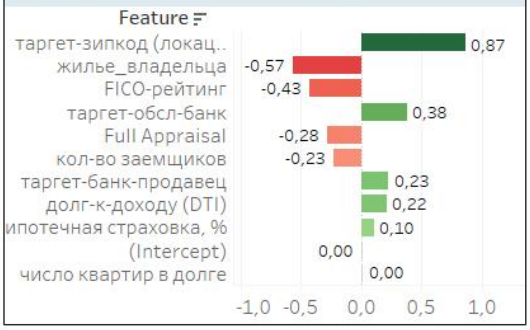


Сравнение моделей по Recall в задаче детектирования дефолтов

- Все модели обучены с оптимизацией Recall
- Порог классификации зафиксирован на 0.5 для честного сравнения
- Минимальный порог (0.0001) дал бы максимальный Recall, но исказил бы оценку
- Географические и институциональные признаки **таргет-энкодированы**, т.к. имеют высокую кардинальность.

Логистическая модель (l1)

Коэффициенты логистической регрессии



Коэффициент отражает изменение логарифма шансов дефолта при увеличении признака на одно стандартное отклонение, т.к. числовые переменные были стандартизированы для корректного сравнения.

Вероятность дефолта повышают: высокая историческая дефолтность по ZIP-коду, значения таргет-признаков для обслуживающего банка и продавца, а также рост DTI и ипотечной страховки.

Снижают риск: проживание владельца, высокий FICO и использование полной оценки недвижимости

Модель случайных лесов

Значимость признаков

Этот случайный лес — ансамбль из 50 деревьев решений с максимальной глубиной 2, обученных на случайных подвыборках признаков (до 10 на каждом сплите). Итоговое решение получается путём агрегации (голосования) всех деревьев, что снижает переобучение и повышает устойчивость.

Модель фокусируется на институциональных и географических паттернах риска (обслуживающий банк, ZIP-код), а также на классических кредитных метриках (DTI, FICO), что подтверждает их значимость в задаче раннего выявления дефолтов

Сравнение моделей по recall

	F1
логистическая	0.443
случайный лес	0.456
XGBoost	0.470

Заключение

Для сравнения была добавлена модель XGBoost, однако несмотря на рост сложности моделей, показатель Recall остался примерно на одном уровне (0.44–0.47). Хотя выбор Recall как основной метрики субъективен, аналогичный вывод наблюдался и при использовании F1-меры и ранговых метрик.

С точки зрения интерпретируемости **предпочтительна логистическая регрессия**: она даёт понятные коэффициенты с направлением влияния признаков. Случайный лес и XGBoost — менее прозрачные модели ("black box"), хотя и более гибкие.

Таким образом, более сложная модель не обязательно лучше классической — особенно в задачах с высокой случайностью и ограниченным количеством полезных признаков. Проект показывает, что успех ML-модели зависит не столько от её сложности, сколько от качества данных и наличия предсказуемых закономерностей.