# Исследование характеристик случайных графов для различения распределений

Куценко Дмитрий, Шатурный Алексей

2025

## 1 Постановка задачи

В данной работе рассматривается задача классификации двух пар параметрических распределений (Laplace$(0, \beta)$ с Normal$(0, \sigma^2)$ и Pareto$(\alpha)$ с Exp$(\lambda)$) с использованием конструкций случайных графов. Основная цель - исследовать, как числовые характеристики графов, построенных на выборках из этих распределений, зависят от параметров распределений и могут быть использованы для построения статистического критерия.

### 1.1 Математическая формулировка

Пусть задана выборка $\hat{\Xi} = (\xi_1, \ldots, \xi_n)$ независимых реализаций случайной величины $\xi$. Требуется проверить две гипотезы:

- $H_0 : \xi \sim \mathcal{N}(0, \sigma^2)$ - нормальное распределение (для Алексея Pareto$(\alpha)$)
- $H_1 : \xi \sim \text{Laplace}(0, \beta)$ - распределение Лапласа (для Алексея Exp$(\lambda)$)

Для решения задачи используются две конструкции случайных графов:

1. **KNN-граф** $\mathcal{GK}(\hat{\Xi}, k)$:
   - Вершины: индексы наблюдений $V = \{1, \ldots, n\}$
   - Рёбра: $(i, j) \in E$ если $\xi_j \in \text{KNN}(\xi_i, k)$ или $\xi_i \in \text{KNN}(\xi_j, k)$
2. **Дистанционный граф** $\mathcal{GD}(\hat{\Xi}, d)$:
   - Вершины: индексы наблюдений $V = \{1, \ldots, n\}$
   - Рёбра: $(i, j) \in E$ если $|\xi_i - \xi_j| \leq d$

## 2 Исследование характеристик графов

В первой части работы исследовалось поведение числовых характеристик графов в зависимости от параметров распределений.

## 2.1 Используемые характеристики

Для анализа были выбраны следующие характеристики графов:

1. Для пары из распределений Normal и Laplace
   - Число треугольников - для KNN-графа
   - Хроматическое число - для дистанционного графа
2. Для пары из распределений Pareto и Exp
   - Число компонент связности - для KNN-графа
   - Размер минимального кликового покрытия - для дистанционного графа

## 2.2 Методология исследования

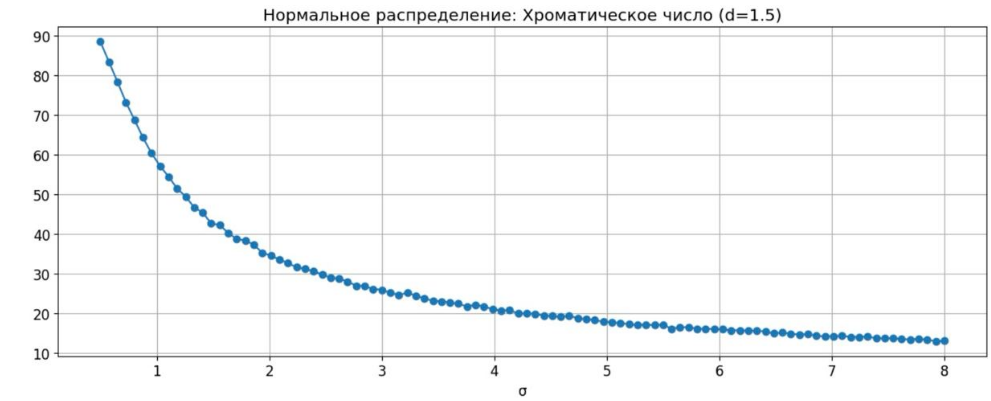
Для каждого типа графа и характеристики проводилось:

1. Фиксация размера выборки $n$ и параметра построения графа ($k$ или $d$)
2. Вариация параметра распределения:
   - Для нормального: $\sigma \in [0.5, 8.0]$
   - Для Лапласа: $\beta \in [0.5, 8.0]$
3. Для каждого набора параметров выполнялось 100 симуляций Монте-Карло
4. Усреднение значений характеристики по симуляциям

## 2.3 Результаты

### 2.3.1 Зависимость характеристик от параметра $\sigma$ нормального распределения

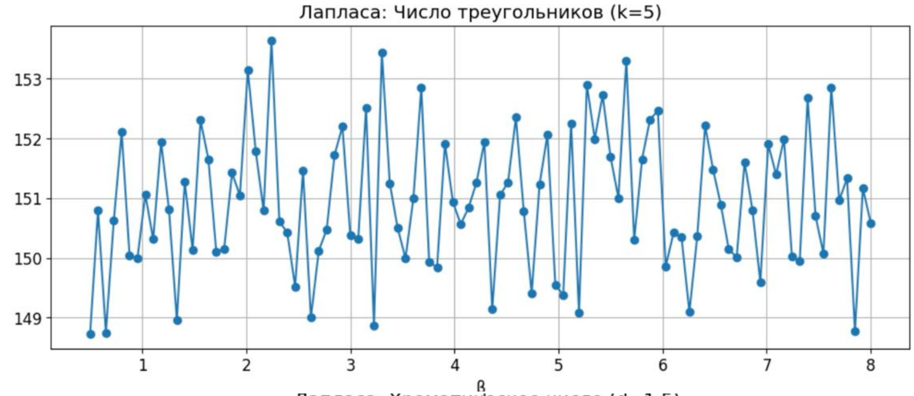
Ниже представлены зависимости характеристик от параметров распределений при фиксированных $n = 100$, $k = 5$, $d = 1.5$.

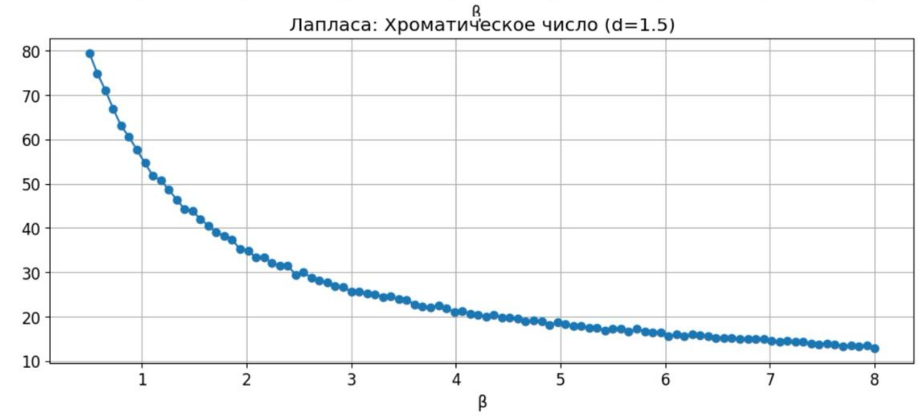




Основные наблюдения:

- **Число треугольников (KNN-граф)**:
  - Тяжело установить явную зависимость числа треугольников в получаемом KNN-графе в зависимости от параметра $\sigma$ нашего распределения. В среднем количество треугольников колеблется около 154
- **Хроматическое число (дистанционный граф)**:
  - Резко убывает с ростом $\sigma$
  - Объяснение: вероятно увеличение разброса приводит к разрежению графа

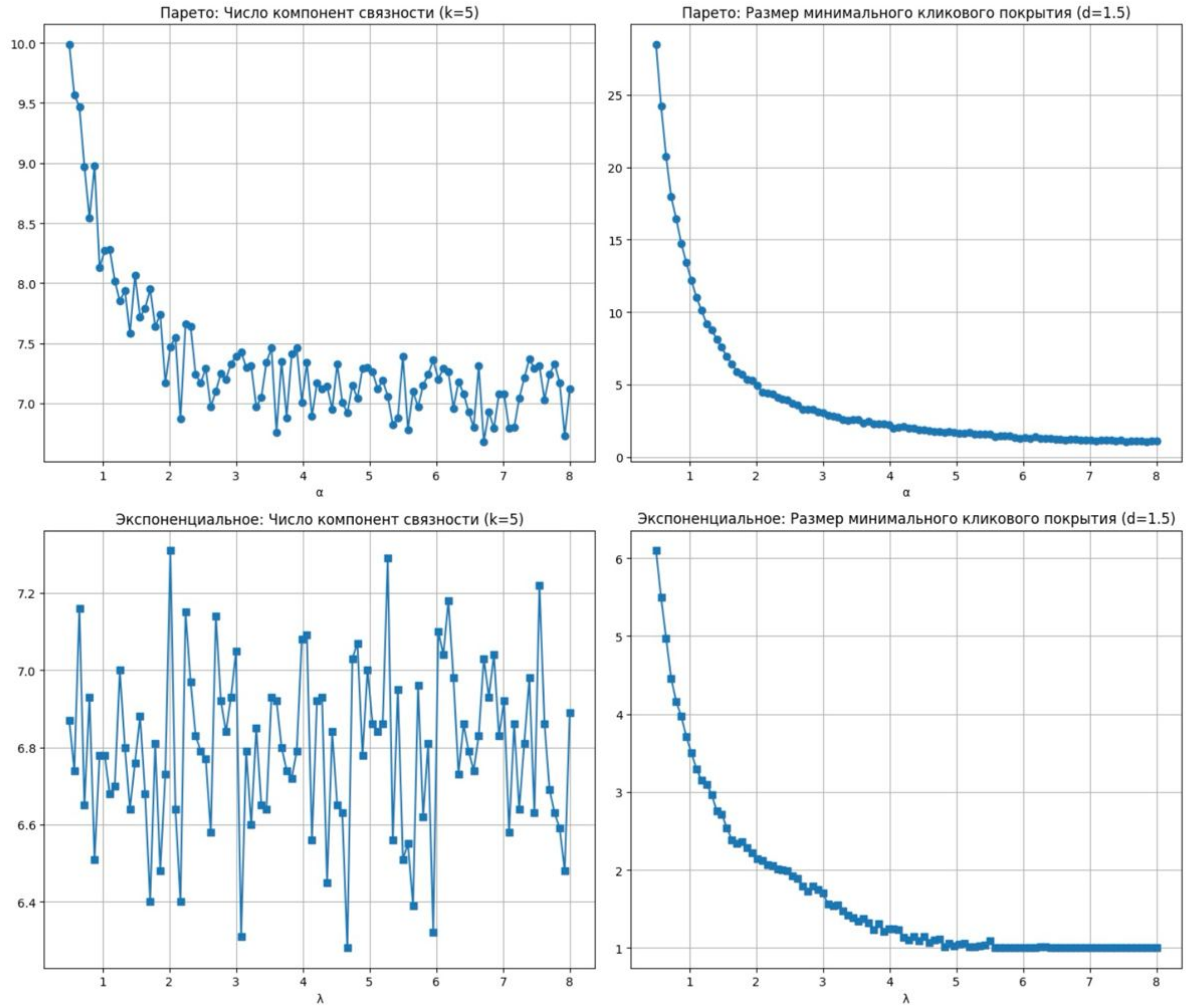
### 2.3.2 Зависимость характеристик от параметра $\beta$ распределения Лапласа




Можем наблюдать ситуацию, похожую на нормальное распределение – видна явная зависимость хроматического числа от параметра $\beta$, в то время как число треугольников в KNN-графе колеблется вокруг значения 151

### 2.3.3 Зависимость характеристик от параметров $\lambda$ экспоненциального распределения и $\alpha$ распределения Парето




Основные наблюдения:

1. Аналогично паре из нормального распределения и распределения Лапласа числовая характеристика дистанционного графа выглядит более информативной и менее шумной

2. С увеличением $\lambda$ и $\alpha$ для обоих распределения размер минимального кликового покрытия дистанционного графа резко падает

### 2.3.4 Исследование поведения числовых характеристик в зависимости от параметров процедуры построения графа и размера выборки при фиксированных параметрах распределений

Будем симулировать выборки при фиксированных параметрах распределений:

1. $\text{Laplace}\left(0, \sqrt{\frac{1}{2}}\right)$

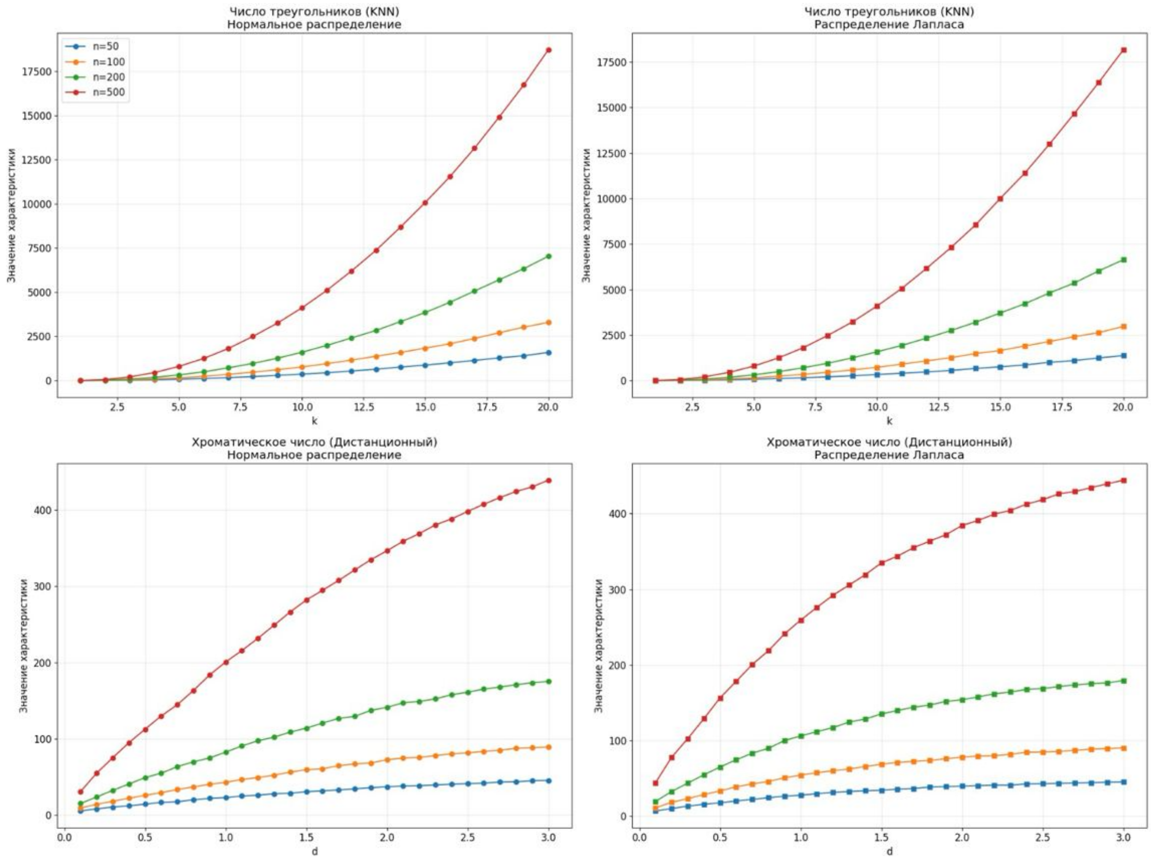
2. $\text{Normal}\,(0, 1)$

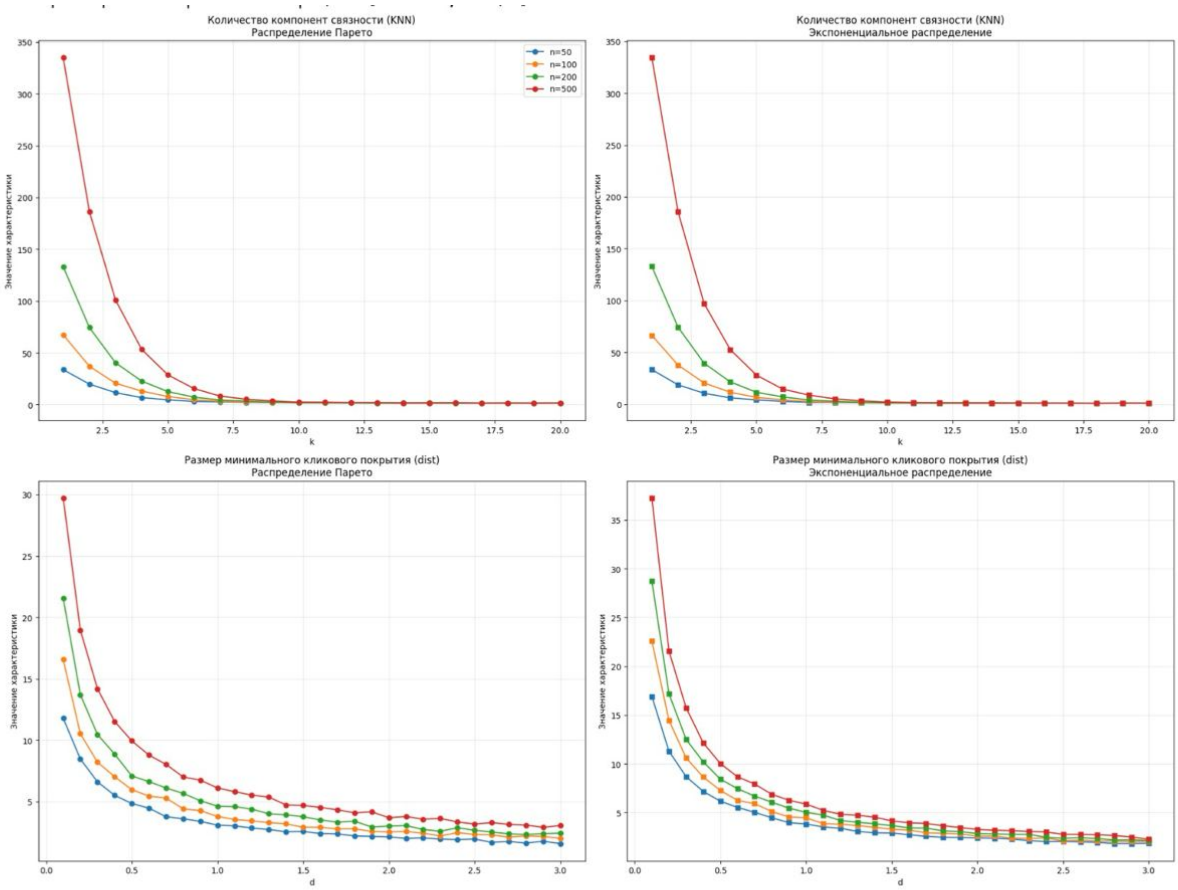
3. $\text{Pareto}(3)$

4. $\text{Exp}\left(\frac{2}{\sqrt{3}}\right)$

Рассмотрим следующие параметры процедуры построения графов

1. $k = 1, 2, 3, ..., 20$
2. $d \in [1, 3]$
3. $n = 50, 100, 200, 500$






Основные наблюдения:

1. При исследовании пары из нормального распределения и распределения Лапласа было установлено, что при увеличении параметров $k$ и $d$ вне зависимости от величины $n$, обе числовые характеристики получаемых случайных графов растут.

2. В паре из распределения Парето и экспоненциального распределения наоборот при увеличении параметров $k$ и $d$ исследуемые числовые характеристики падали

3. В обоих случаях логично с увеличением размера выборок (то есть параметра $n$) значение числовых характеристик росло, но характер роста (или наоборот падения) оставался прежним

### 2.3.5 Построение критических областей

# 3 Заключение первой части

Проведенное исследование показало:

1. Числовые характеристики случайных графов чувствительны к параметрам распределений

2. Пронаблюдали род зависимости каждой из выбранных числовых характеристик KNN-графа и дистанционного графа в зависимости от различных параметров распределений

3. Более информативную и явную зависимость удалось выявить при исследовании дистанционных графов

- Для пары Laplace и Normal хорошо показало себя хроматическое число
- Для пары Pareto и Exp хорошо показал себя размер минимального кликового покрытия

# 4 Применение нескольких характеристик для проверки гипотезы

### 4.0.1 Выбор модели случайного графа

По итогам исследования поведения числовых характеристик при изменении параметров распределений и параметров процедуры построения графа было принято решение работать именно с дистанционным графом, так как числовые характеристики дистанционного графа проявляли себя как более информативные.

### 4.0.2 Фиксирование параметров распределений и параметров процедуры построения графа

1. $\text{Laplace}\left(0, \sqrt{\frac{1}{2}}\right)$
2. $\text{Normal}\,(0, 1)$
3. Pareto(3)
4. $\text{Exp}\left(\frac{2}{\sqrt{3}}\right)$
5. $d = 1.5$

### 4.0.3 Построение обучающего набора данных

1. Построим набор данных из 5000 объектов для $n = 25$ (В таком наборе данных поровну объектов для каждого из распределений в паре – по 2500)
2. Набор данных из 5000 объектов для $n = 100$
3. Набор данных из 200 объектов для $n = 500$

### 4.0.4 Выбор моделей машинного обучения

Для классификации распределений для обоих пар использовались 3 классических алгоритма классификации, позволяющих интерпретировать важность каждого из признаков:

1. Логистическая регрессия
2. Случайный лес
3. Градиентный бустинг

#### 4.0.5 Оценка метрик качества классификации нормального распределения и распределения Лапласа

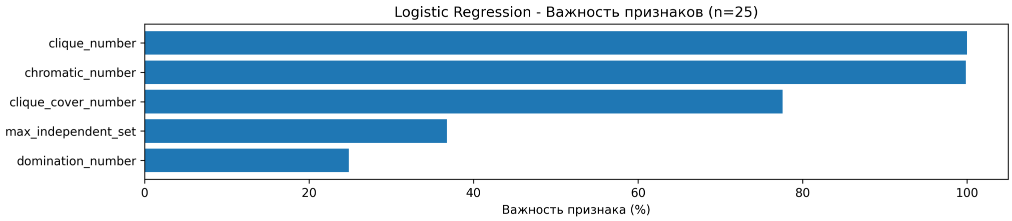
```
==================================================
Анализ для размера выборки n = 25
==================================================

Результаты для n=25:
                 Model  Size  Accuracy  Precision  Recall        F1
0  Logistic Regression    25     0.759   0.751456   0.774  0.762562
1        Random Forest    25     0.774   0.795259   0.738  0.765560
2             CatBoost    25     0.775   0.796976   0.738  0.766355

==================================================
Анализ для размера выборки n = 100
==================================================

Результаты для n=100:
                 Model  Size  Accuracy  Precision  Recall        F1
0  Logistic Regression   100     0.911   0.915152   0.906  0.910553
1        Random Forest   100     0.909   0.914807   0.902  0.908359
2             CatBoost   100     0.907   0.904573   0.910  0.907278

==================================================
Анализ для размера выборки n = 500
==================================================

Результаты для n=500:
                 Model  Size  Accuracy  Precision  Recall        F1
0  Logistic Regression   500      0.95        1.0     0.9  0.947368
1        Random Forest   500      1.00        1.0     1.0  1.000000
2             CatBoost   500      1.00        1.0     1.0  1.000000
```

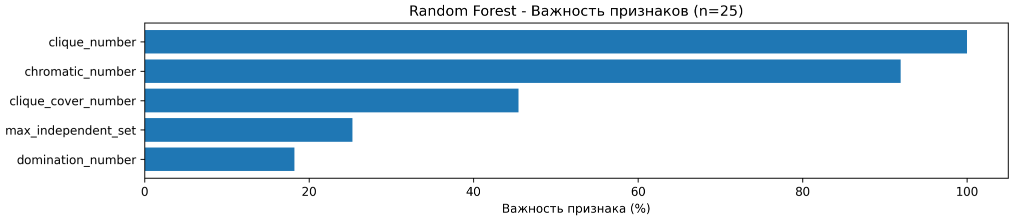
#### 4.0.6 Оценка метрик качества классификации экспоненциального распределения и распределения Парето

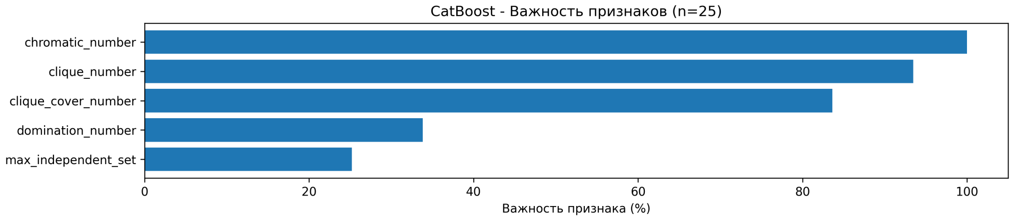
```
==================================================
Анализ для размера выборки n = 25
==================================================

Результаты для n=25:
                 Model  Size  Accuracy  Precision  Recall        F1
0  Logistic Regression    25     0.814   0.829832   0.790  0.809426
1        Random Forest    25     0.814   0.835470   0.782  0.807851
2             CatBoost    25     0.815   0.835821   0.784  0.809082

==================================================
Анализ для размера выборки n = 100
==================================================

Результаты для n=100:
                 Model  Size  Accuracy  Precision  Recall        F1
0  Logistic Regression   100     0.977   0.979879   0.974  0.976931
1        Random Forest   100     0.977   0.979879   0.974  0.976931
2             CatBoost   100     0.977   0.979879   0.974  0.976931

==================================================
Анализ для размера выборки n = 500
==================================================

Результаты для n=500:
                 Model  Size  Accuracy  Precision  Recall   F1
0  Logistic Regression   500       1.0        1.0     1.0  1.0
1        Random Forest   500       1.0        1.0     1.0  1.0
2             CatBoost   500       1.0        1.0     1.0  1.0
```

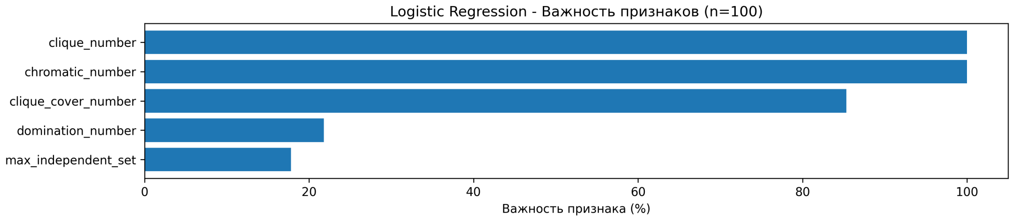
Можем видеть, что при увеличении параметра $n$ – размера выборок, метрики качества классификации растут. Выбранные модели очень хорошо справляются с поставленной задачей

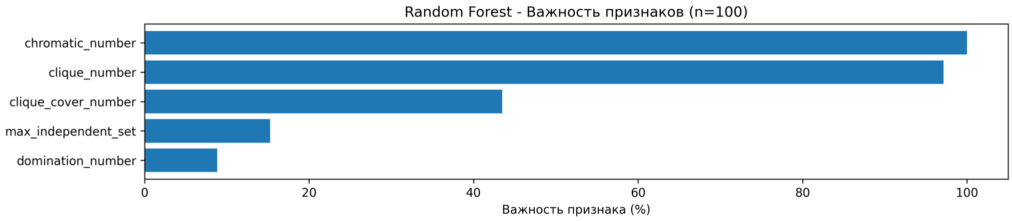
### 4.0.7 Исследование важности характеристик, как признаков классификации нормального распределения и распределения Лапласа

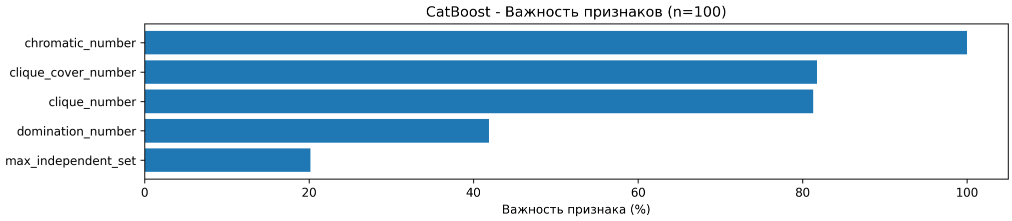


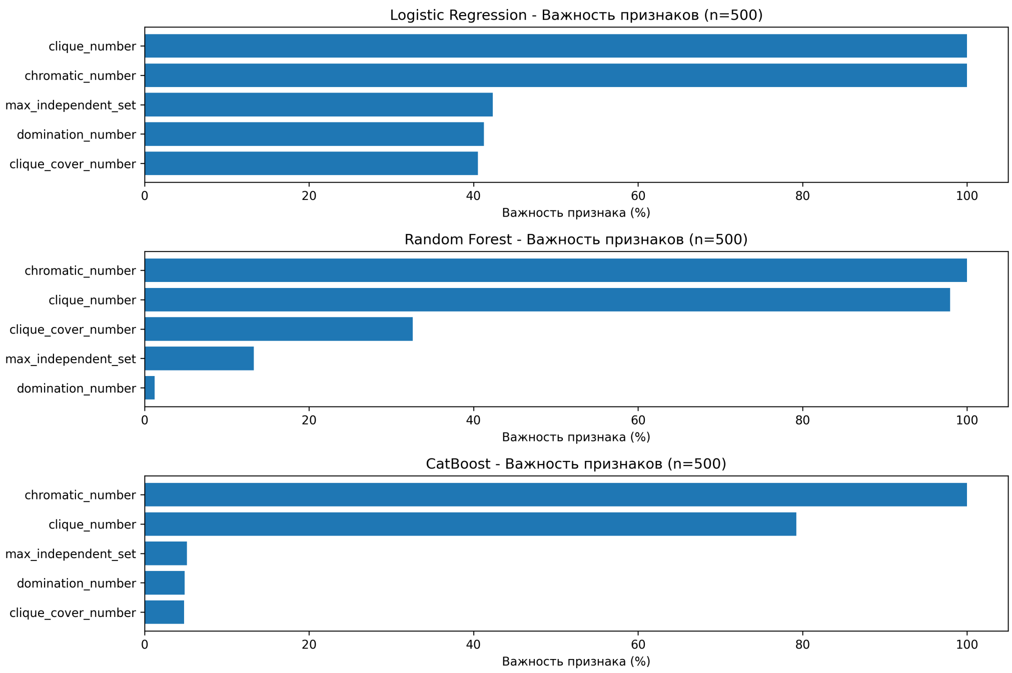






Как мы можем видеть для каждого из наших алгоритмов самыми важными признаками оказались кликовое число и хроматическое число. Также при изменении величины $n$ можем заметить небольшие изменения важности остальных признаков.

#### 4.0.8 Выводы о вероятности ошибки первого рода и мощности полученных статистических критериев для классификации нормального распределения и распределения Лапласа

```
==================================================
Статистический анализ для размера выборки n = 25
==================================================

Статистические метрики для n=25:
                 Model  Type I Error  Power
0  Logistic Regression         0.256  0.774
1        Random Forest         0.190  0.738
2             CatBoost         0.188  0.738

==================================================
Статистический анализ для размера выборки n = 100
==================================================

Статистические метрики для n=100:
                 Model  Type I Error  Power
0  Logistic Regression         0.084  0.906
1        Random Forest         0.084  0.902
2             CatBoost         0.096  0.910

==================================================
Статистический анализ для размера выборки n = 500
==================================================

Статистические метрики для n=500:
                 Model  Type I Error  Power
0  Logistic Regression           0.0    0.9
1        Random Forest           0.0    1.0
2             CatBoost           0.0    1.0
```

Получили хорошие значения мощности построенных критериев и довольно низкие вероятности ошибки первого рода

#### 4.0.9 Выводы о вероятности ошибки первого рода и мощности полученных статистических критериев для классификации экспоненциального распределения и распределения Парето

```
==================================================
Статистический анализ для размера выборки n = 25
==================================================

Статистические метрики для n=25:
                 Model  Type I Error  Power
0  Logistic Regression         0.162  0.790
1        Random Forest         0.154  0.782
2             CatBoost         0.154  0.784

==================================================
Статистический анализ для размера выборки n = 100
==================================================

Статистические метрики для n=100:
                 Model  Type I Error  Power
0  Logistic Regression          0.02  0.974
1        Random Forest          0.02  0.974
2             CatBoost          0.02  0.974

==================================================
Статистический анализ для размера выборки n = 500
==================================================

Статистические метрики для n=500:
                 Model  Type I Error  Power
0  Logistic Regression           0.0    1.0
1        Random Forest           0.0    1.0
2             CatBoost           0.0    1.0
```

Общие выводы:

1. Для второй пары распределений мощность полученных критериев оказалась еще лучше, что делает построенные статистические критерии еще более эффективными, чем построенный в первой части.

2. Можем заметить, что при увеличении $n$ логично росла мощность каждого из критериев и падала вероятность ошибки первого рода