

Ибаева Анастасия
Смотрова Кристина

Руководитель:
Мария Воробьева

18 июня 2022 года

Построение модели прогнозирования страхового тарифа



Команда

Анастасия



Кристина



Задачи

[https://github.com/anastasia
ibaeva/DABproject](https://github.com/anastasiaibaeva/DABproject)

1

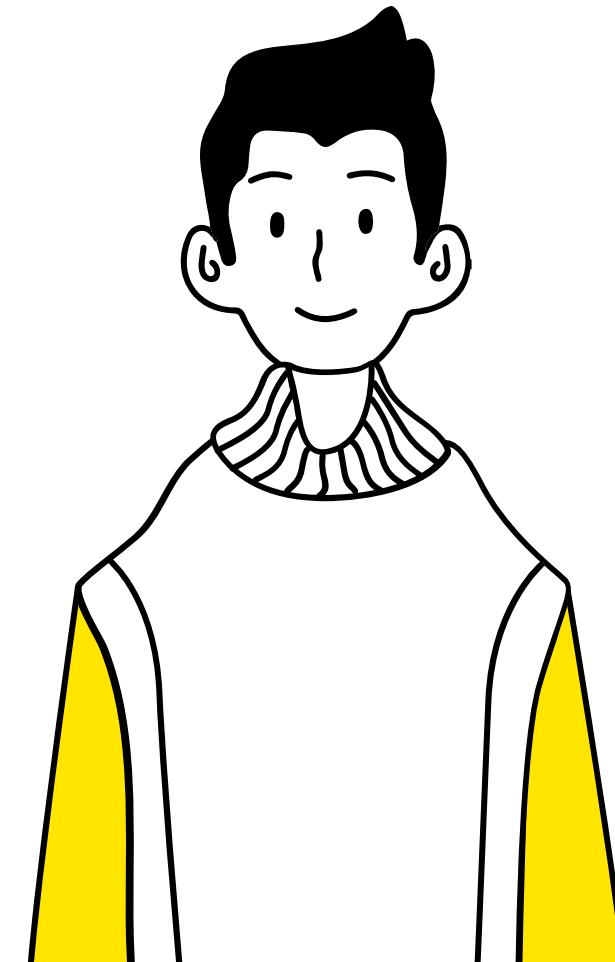
Анализ предоставленных данных,
создание производных переменных

3

Выбор наилучшей модели

2

Разработка моделей



Этапы проекта



Анастасия

1

Выделение таргета

Кристина

2

Предобработка данных:
заполнение пропусков,
удаление лишней
информации

Анастасия

3

Выделение новых признаков

вместе

4

Анализ моделей

5

Построение моделей

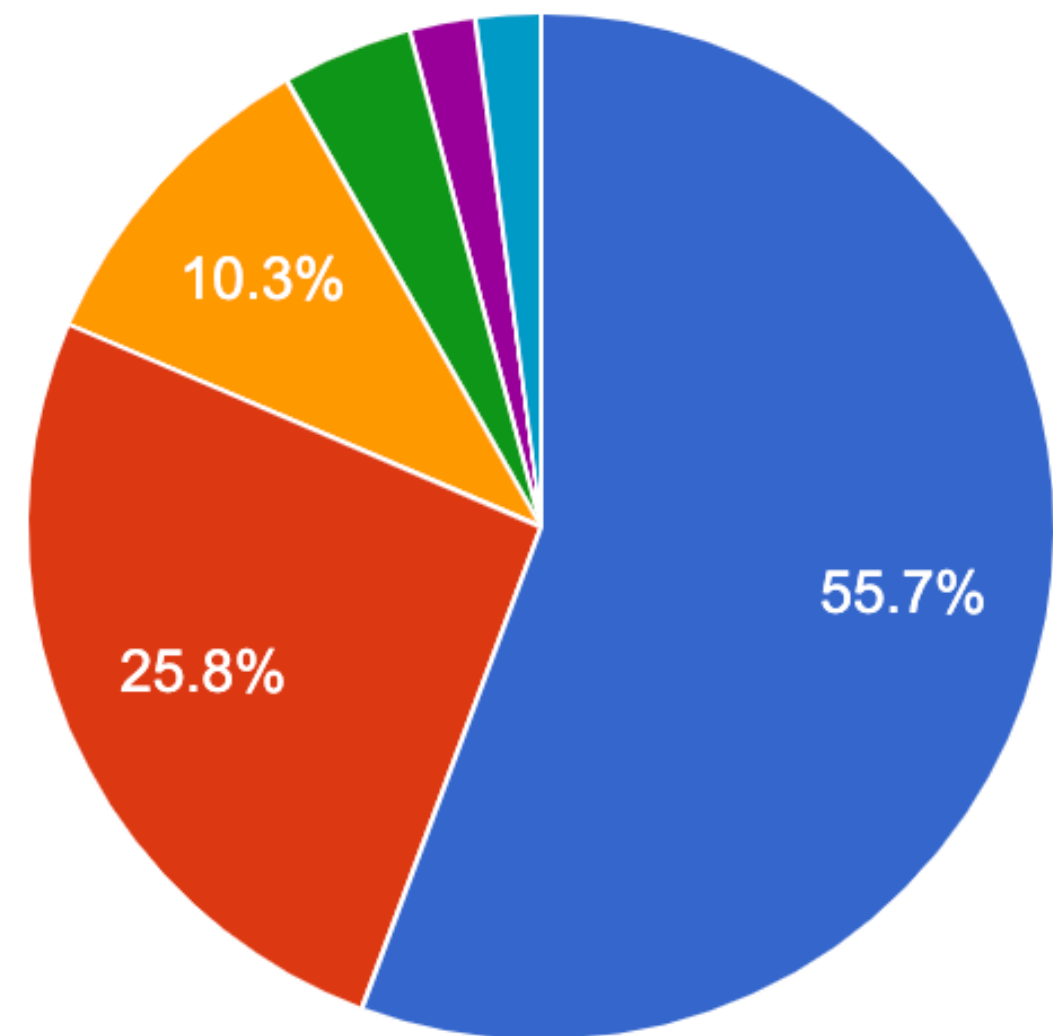
Кристина

6

Анализ результатов

Таргет

Сумма убытков, которые
принес клиент страховой за
месяц



Работа с данными

13 категориальных и
10 числовых признаков

547868 объектов, 24 признака

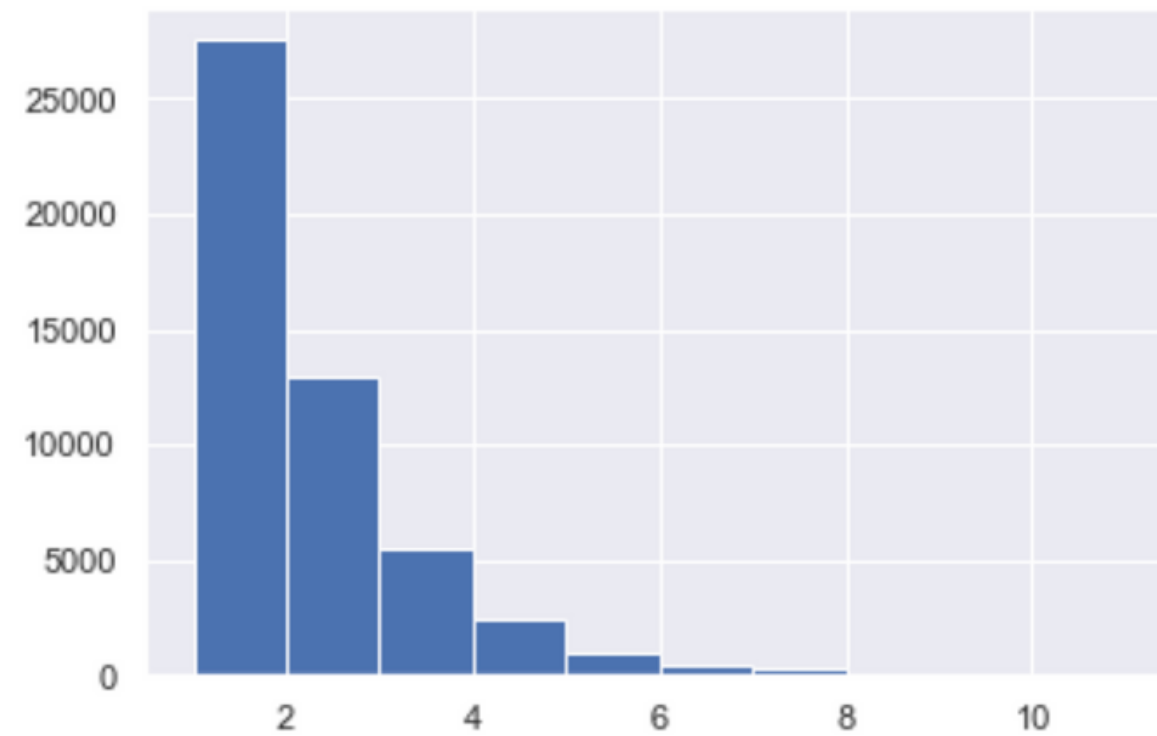


199808 объектов, 23 признака

- Заполнение пропусков
- Удаление признаков: patient_id, folk_4_code, has_doctor, is_paid_by_patient
- Приведение признака date к единому формату
- Анализ распределения данных по признакам

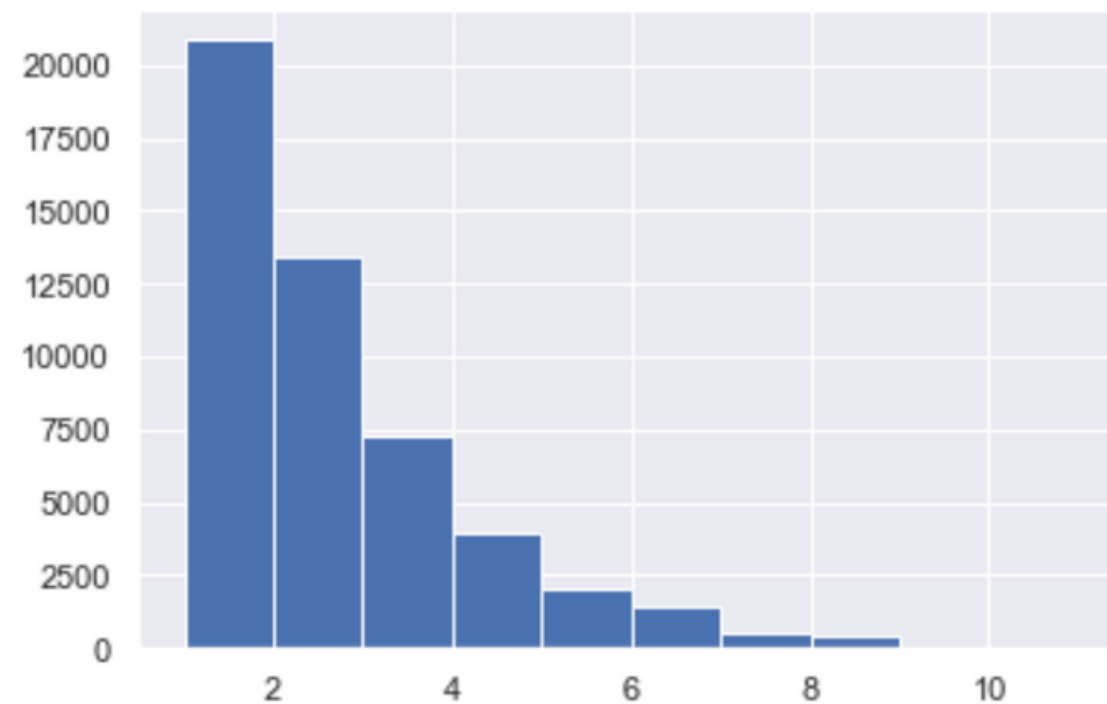
Новые признаки

1



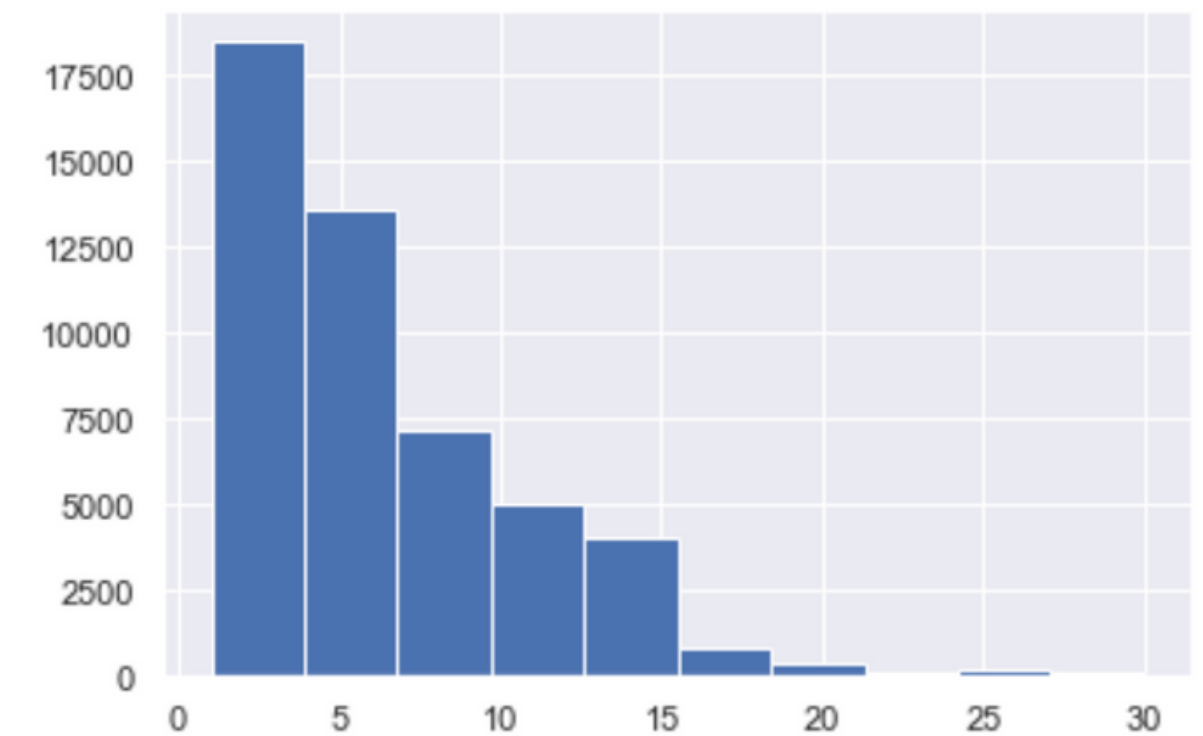
Количество страховых писем за неделю

2



Количество страховых писем за месяц

3



Количество назначенных клиенту услуг за один прием

Decision tree

Строит модели регрессии или классификации в виде древовидной структуры

Random forest

Алгоритм обучения с учителем, который использует метод ансамблевого обучения для регрессии.

CatBoost

Библиотека, использующая градиентный бустинг

Используемые модели

Метрики

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$R^2 = 1 - \frac{\frac{1}{n} \sum_i |y_i - \hat{y}_i|^2}{\frac{1}{n} \sum_i |y_i - \bar{y}_{train}|^2},$$



Подбор гиперпараметров

Random forest

max_features: ['auto', 'log2', 'sqrt']

max_depth: [2, 4, 6, 8, 10, 20, 50]

min_samples_leaf: [1, 4, 8, 32]

Decision tree

max_features: ['auto', 'log2', 'sqrt']

max_depth: [2, 4, 6, 8, 10, 20, 50]

min_samples_leaf: [1, 4, 8, 32]

CatBoost

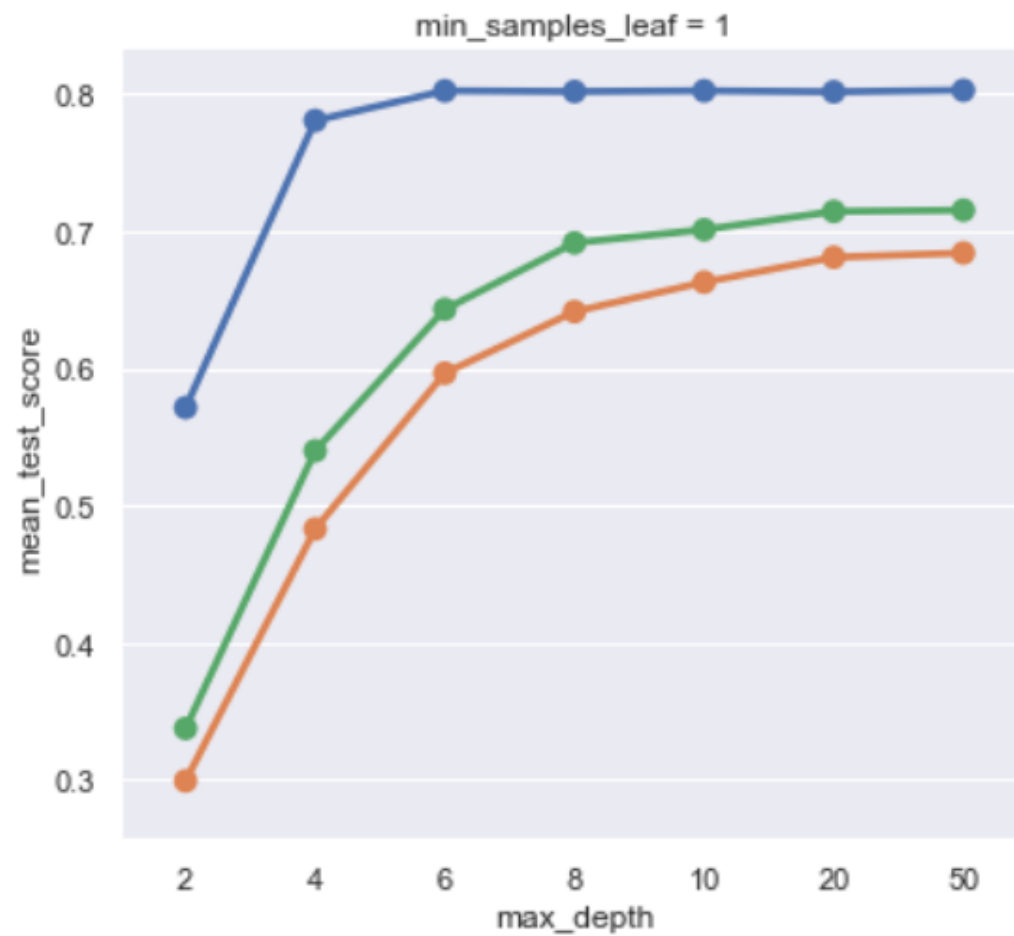
depth: [1, 3, 5, 7, 9]

iterations: [1, 11, 21, 31, 41, 51, 61, 71, 81, 91]

learning_rate: [0.001, 0.01, 0.02, 0.03, 0.5, 0.1, 1]

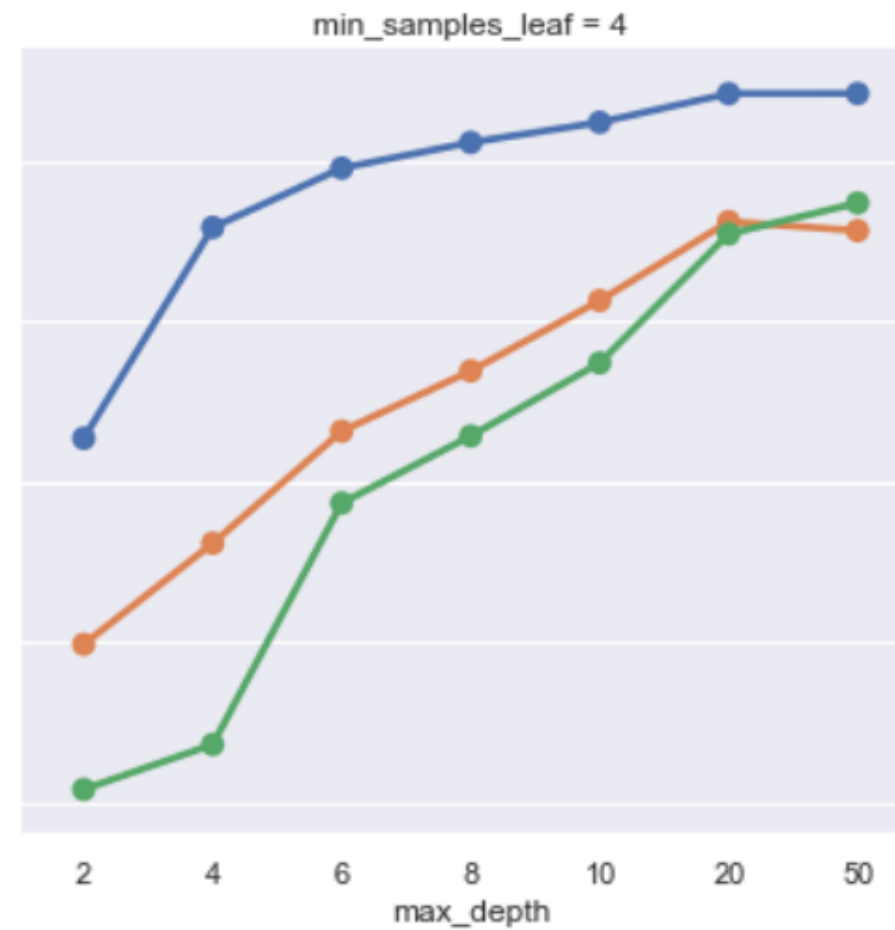
Оптимальные гиперпараметры

Random forest



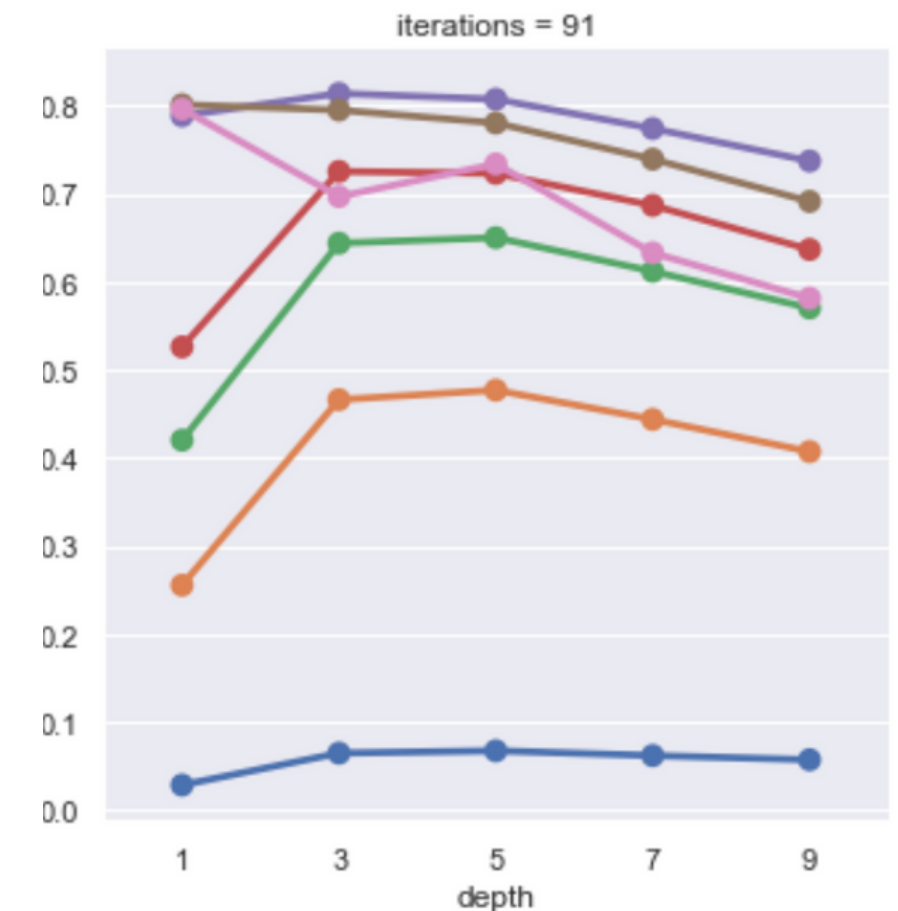
'max_depth': 50, 'max_features':
'auto', 'min_samples_leaf': 1

Decision tree



'max_depth': 50, 'max_features':
'auto', 'min_samples_leaf': 4

CatBoost



'depth': 3, 'iterations': 91,
'learning_rate': 0.1

Полученные результаты

	Decision tree	Random forest	CatBoost
RMSE	3.804142	3.03141	5.851731
MAE	0.589576	0.743685	2.525766
R2	0.902283	0.937949	0.768779

Спасибо за внимание!

