Предсказание спам заявок на услуги по заполненным полям формы заявки

май, 2023

Клочнева Е.Ю.
GB

Содержание

- Обзор данных
- Анализ проблемы
- Цель проекта
- Алгоритм исследования и построения модели
- Процесс реализации
- Выводы

Обзор данных

Датасет включает в себя данные, которые пользователь заполняет при оформлении заявки на подключение услуги. Также учтено время оформления заявки.
На данных будем строить модель, которая сможет определять сразу, является ли данная заявка “рабочей” или это спам и боты.

Поле	Значение	Поле	Значение
reject_reason	Целевая переменная	email	Контактный емейл
priority	Приоритет заявки (по умолчанию 2)	industry	Сфера деятельности
lead_source	Источник заявки (сайт, звонок и т. д.)	web_form_of_consent	Согласие на подписку
description	Описание (комментарий)	marketing_events	Маркетинговое мероприятие
product	Интересующий продукт	status	Статус заявки
phone	Контактный телефон	created_at	Время создание заявки

Цель проекта

На этапе поступления заявки определять, стоит ли тратить стандартное время на ее обработку.	Уменьшить количество нагрузки на операторов и менеджеров.	Выявить “слабые стороны” при оформлении заявки через форму на сайте
Предложить возможные меры и методы для уменьшения спам заявок.	Определить критерии, позволяющие с определенной долей вероятности определять “рабочие” заявки	Увеличить скорость обработки “рабочей заявки”

Алгоритм исследования и построения модели

- **Анализ полноты и типов данных**
- **Преобразование признаков**
- **Построение нескольких моделей, для выбора наилучшей**
- **Метрики качества**
- **Оценка влияния признаков**

What we Do

Add your own subtitle here.

Aliquam eget tincidunt ligula. Quisque eget magna non diam blandit cursus. Ut euismod turpis nisl, vitae dictum nibh posuere in. In eget magna sit amet metus dictum placerat.

Nam eu faucibus enim, id ullamcorper tellus. Duis eget orci ac ipsum varius ultricies. Nam mattis et nisl quis fermentum. Cras eget mollis sapien. Integer fringilla sit amet est eget aliquet. Phasellus tincidunt metus ac mattis fringilla.

Nam in viverra felis, et ullamcorper ipsum. Fusce condimentum eget massa non malesuada. Integer et facilisis est. Mauris interdum erat turpis, at consectetur tellus viverra sed.

Nam eu faucibus enim, id ullamcorper tellus. Duis eget orci ac ipsum varius ultricies. Nam mattis et nisl quis fermentum. Cras eget mollis sapien. Integer fringilla sit amet est eget aliquet. Phasellus tincidunt metus ac mattis fringilla. Nam in viverra felis, et ullamcorper ipsum. Fusce condimentum eget massa non malesuada.

This is a description

This is a description

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

This is a description

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

This is a description

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

This is a description

Lorem ipsum dolor sit amet, consectetur adipiscing elit.

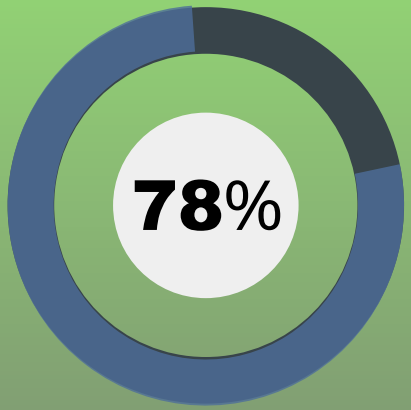
Our Data

Add your own subtitle here.

	Column 1	Column 2	Column 3	Column 4
Row 1	This is your text.	This is your text.	This is your text.	This is your text.
Row 2	This is your text.	This is your text.	This is your text.	This is your text.
Row 3	This is your text.	This is your text.	This is your text.	This is your text.
Row 4	This is your text.	This is your text.	This is your text.	This is your text.
Row 5	This is your text.	This is your text.	This is your text.	This is your text.

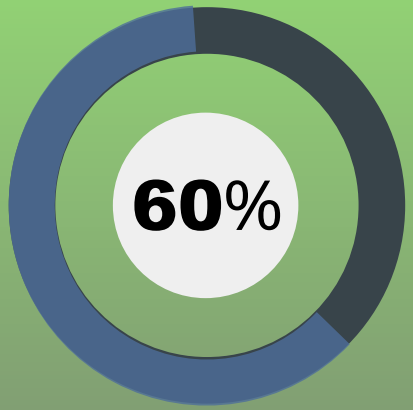
Donut Charts

Add your own subtitle here.



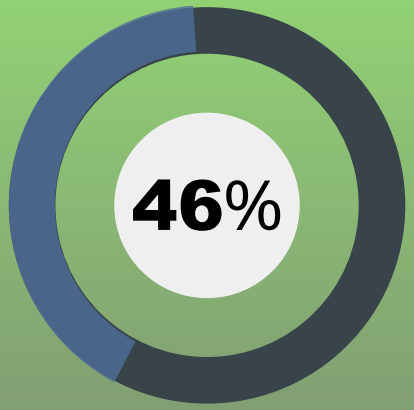
$44,123

This is a Metric



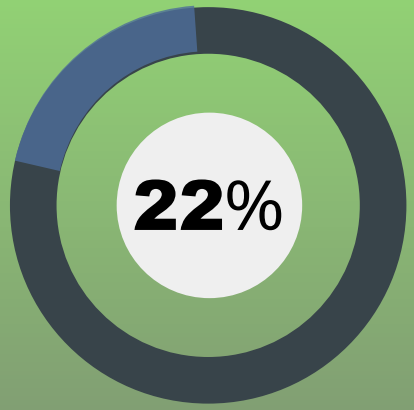
$44,123

This is a Metric



$44,123

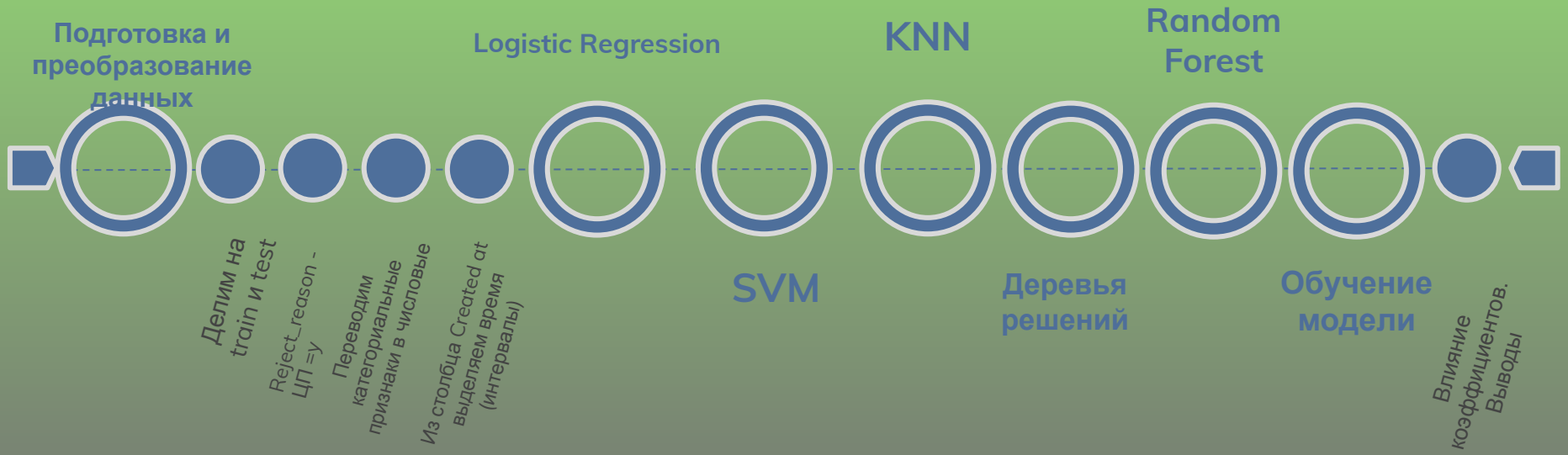
This is a Metric



$44,123

This is a Metric

Процесс реализации



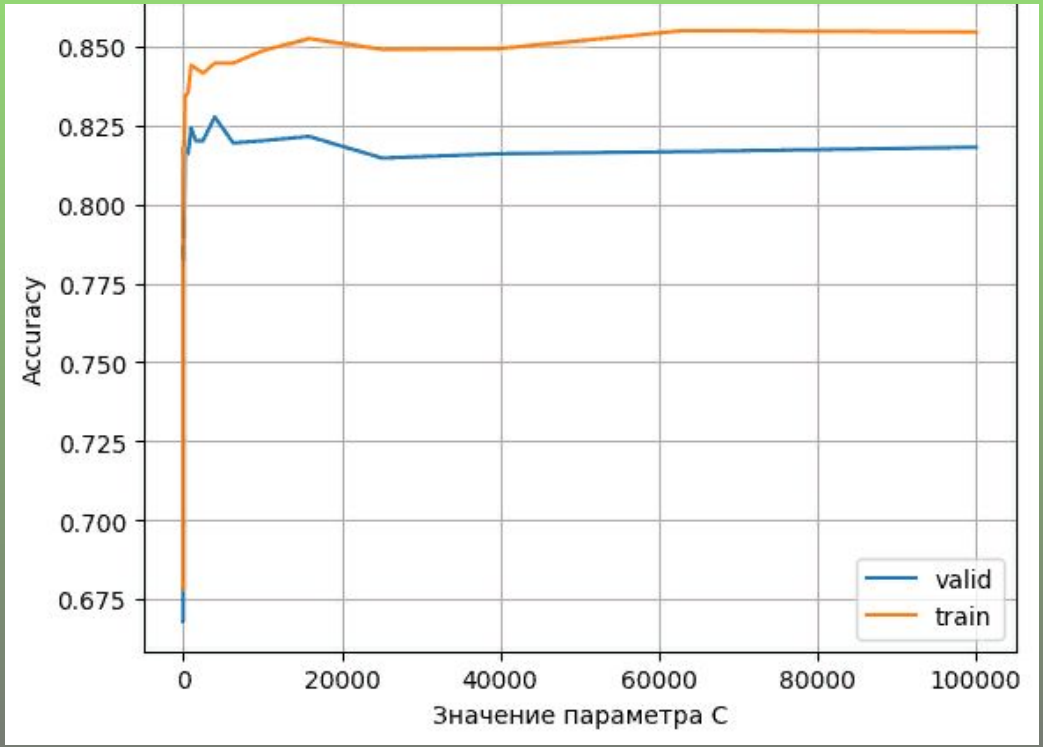
Logistic Regression

Показала достаточно низкую точность

```
0.7886904761904762
```

```
array([[0.96659347, 0.03340653],
       [0.05990893, 0.94009107],
       [0.95624232, 0.04375768],
       [0.99799447, 0.00200553],
       [0.0691506 3, 0.93084937],
       [0.99566825, 0.00433175],
       [0.2499684 , 0.7500316 ],
       [0.31364821, 0.68635179],
       [0.55498925, 0.44501075],
       [0.99685963, 0.00314037]])
```

SVM (Support Vector Machine)

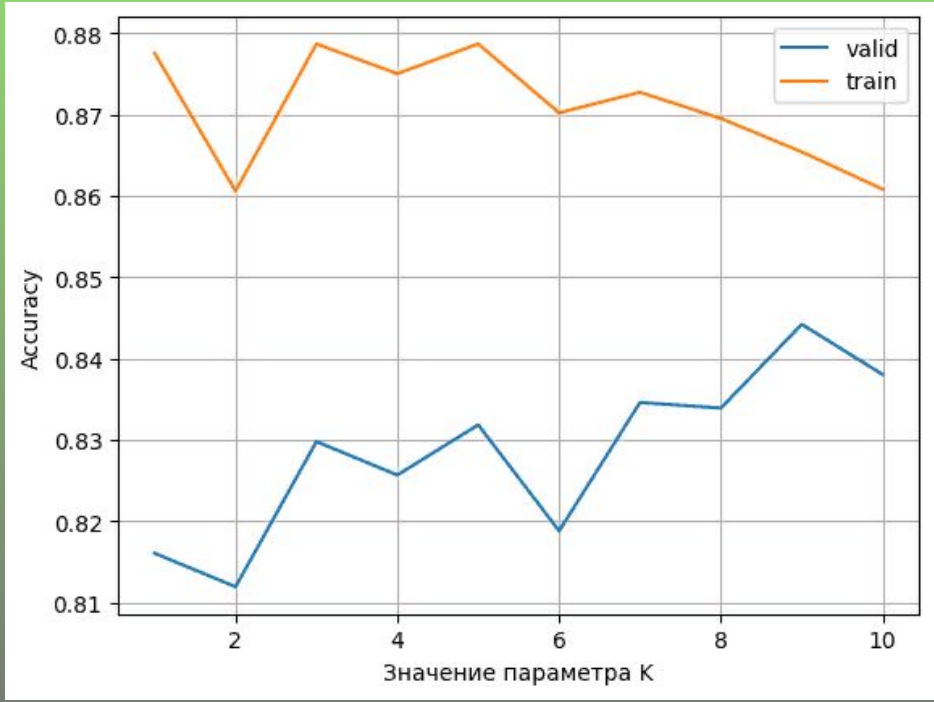


Точность на валидационных данных до определённого момента растёт, но затем начинает падать. Пик точности попадает приблизительно на значение C=4500 . Дальше мы видим спад точности на валидационных данных, однако, точность на тренировочных данных продолжает расти. Это означает, что модель начинает переобучаться.

Можно заметить, что точность на валидационных данных при C=4500 уже выше, чем при использовании логистической регрессии, а также выше, чем при использовании SVM без настройки параметров.

KNN (K Nearest Neighbours)

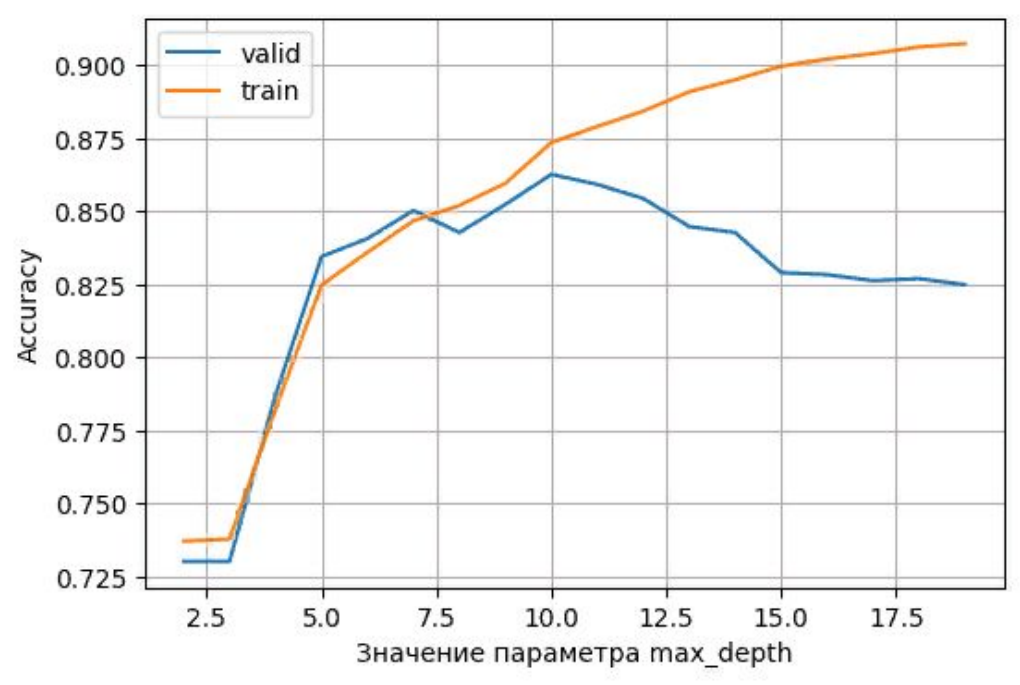
Оптимальным значением является k=7



При к=7 получаем

```
0.8345916266300618
```

Деревья решений



При достаточно больших значениях параметра `max_depth` точность на тренировочных данных почти достигает 1. На валидационных же данных точность достигает своего пика приблизительно на значении `max_depth = 10`, а затем начинает падать.

```
0.8627316403568978
```

Random Forest

Лучшие параметры

```
{'max_depth': 9, 'max_features': 8, 'n_estimators': 200}
```

```
0.8551818805765271
```

```
Лучший показатель метрики качества
```

Влияние признаков

Add your own subtitle here.

Источник лида

Время оформления заявки

День оформления заявки

Указние емейл адреса

Источник лида

Выводы

Как мы можем увидеть, наибольшее влияние на то, является ли заявка спамом или “рабочей”, оказывает признак ИСТОЧНИКА ЛИДА (САЙТ). В действительности так и есть. С Сайта приходит наибольшее количество спама. Интересно было подтвердить интуитивную догадку о том, что время также является признаком, который оказывает влияние на нашу ЦА.

В качестве модели выбор был остановлен на Random Forest, как на наиболее точной по показателю качества.

Учитывая полученные данные можно будет с высокой вероятностью, сразу отделять “рабочие заявки” от спама. Спам не будет отбрасываться, т.к. Вероятность не 100%, но будет обрабатываться по короткому сценарию, для экономии времени и трудозатрат.

“Спама бояться - соб@ку не заводить”