Необходимо угадать эмоции посетителей пространства «Точки кипения». Дана таблица с данными по посетителям за несколько последних дней (эмоции распознаются по изображениям, полученным с камер при входе в пространство). Программа должна предсказать настроение участников, которые придут в «Точку кипения» 12 мая (данные фиксируются через систему распознавания лиц на стойке входа).

**Входные данные**: таблица статистики по настроению посетителей посещения московской точки. Доступно семь эмоций: злость, радость, грусть, удивление, отвращение, страх, безразличие. Каждая эмоция выражается целым числом (процентом посетителей). Например, грусть – 50%, радость – 30%, удивление – 20%, остальные значения = 0.

Для решения задачи были собраны данные по мероприятиям с сайта leader-id.ru (тип, время проведения, длительность, организатор), погоде (температура воздуха днем/ночью, наличие осадков, облачность, атмосферное давление), общие данные (день недели).

**Проблемы:**

1. Небольшой набор наблюдений (23 дня). В выборке возможно появление выбросов (нестандартных наблюдений), в маленьком наборе данных их влияние может очень сильно искажать картину.
2. Распределение эмоций дано не по каждому мероприятию или конкретному времени, а по всему дню, при этом в один день может проходить около 10 мероприятий, так что нельзя точно отследить зависимость эмоций от типа мероприятия (семинар, тренинг, лекция, форсайт и т.п.), организатора, времени проведения и т.д. Из-за этого был использован подсчет различных мероприятий за день. Например, если количество конференций в нескольких наблюдениях заметно превышает количество заседаний рабочих групп и в эти дни фиксируется снижение уровня грусти, возможно, люди реже грустят, когда идут на конференцию.
3. Входными данными служат не эмоции, которые испытывает человек, а то, как программа распознала его мимику. Например, печаль обычно распознается при немного опущенных верхних веках и опущенных уголках рта, что может быть следствием сонливости, а удивление и страх определяются похожим образом (приподнятые брови, широко раскрытые глаза). При этом точный алгоритм, по которому происходило распознавание, неизвестен.
4. Вероятно, на сайте leader-id указываются не все мероприятия, которые проходят в Точке кипения, так что отследить истинную зависимость настроения от расписания сложно.

**Признаки для анализа:**

|  |  |  |
| --- | --- | --- |
| Название | Тип признака | Описание |
| data | порядковый | дата в формате «день-месяц» |
| emo | номинальный | название эмоции  (angry, disgust, fear, happy, neutral, sad, surprise) |
| emo value | количественный | процент посетителей за день, испытывающих данную эмоцию (0 – 100 %) |
| week day | порядковый | день недели (1 – 7) |
| avg time | количественный | среднее за день время начала мероприятий |
| morning | количественный | количество утренних мероприятий за день (начало до 11:00) |
| day | количественный | количество дневных мероприятий за день (начало с 11:00 до 16:00) |
| evening | количественный | количество вечерних мероприятий за день (начало после 16:00) |
| half-1 | количественный | количество мероприятий в первой половине дня  (начало до 14:00) |
| half-2 | количественный | количество мероприятий во второй половине дня  (начало после 14:00) |
| d temp | количественный | средняя за день дневная температура |
| n temp | количественный | средняя за день ночная температура |
| avg temp | количественный | среднесуточная температура |
| cloud | порядковый | облачность в текущий день (по шкале от 0 до 3, где 0 – ясно, 3 – пасмурно) |
| rain | бинарный | дождь в текущий день (был/не было) |
| pressure | количественный | атмосферное давление в текущий день |
| meeting | количественный | количество совещаний за день |
| sem-train | количественный | количество семинаров/тренингов за день |
| lection | количественный | количество лекций за день |
| bus-accelerate | количественный | количество бизнес-акселераторов за день |
| strat-session | количественный | количество стратегических сессий за день |
| conference | количественный | количество конференций за день |
| work-group | количественный | количество встреч рабочих групп за день |
| celebration | количественный | количество праздников за день |
| forum | количественный | количество форумов за день |
| round-table | количественный | количество круглых столов за день |
| foresight | количественный | количество форсайтов за день |
| competition | количественный | количество конкурсов за день |
| open | количественный | количество открытых (может попасть любой желающий) мероприятий за день |
| close | количественный | количество закрытых мероприятий за день |
| nti | бинарный | наличие мероприятий Университета НТИ  в текущий день |

**Анализ данных**

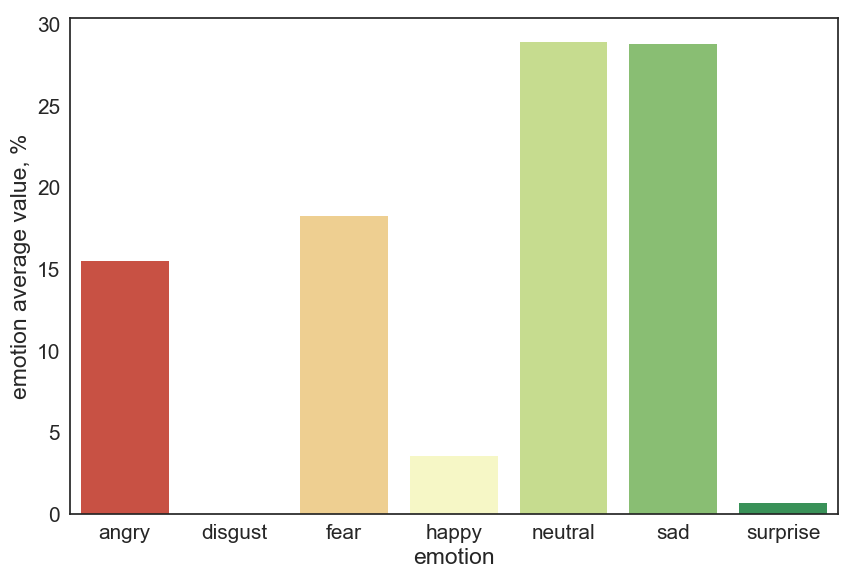
Согласно программе распознавания эмоций, люди, приходящие в «Точку кипения», в основном, не испытывают эмоций или испытывают грусть, страх или раздражение.

В таблице приведены средние, минимальные и максимальные значения эмоций (в порядке убывания среднего), разброс значений:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| эмоция | среднее  значение, % | min, % | max, % | разброс |
| neutral | 28.95 | 0 | 93.85 | 93.85 |
| sad | 28.82 | 0 | 100 | 100 |
| fear | 18,26 | 0 | 100 | 100 |
| angry | 15,53 | 0 | 90.54 | 90.54 |
| happy | 3.57 | 0 | 38.1 | 38.1 |
| surprise | 0.71 | 0 | 5.77 | 5.77 |
| disgust | 0 | 0 | 0 | 0 |

Наиболее стабильны отвращение (полностью отсутствует), удивление (абсолютная разница между максимумом и минимумом – 5,77 %), радость (разброс 38 %).

Средние по выборке значения эмоций приведены также на графике:



Для каждой эмоции найдены коэффициенты линейной корреляции, от -1 до 1, где:

* 1 означает 100%-ю прямую связь, при возрастании одного признака второй также возрастает, можно определить величину изменения со 100%-1 точностью;
* -1 – 100%-обратная связь, при возрастании одного признака второй убывает;
* 0 – отсутствие связи; чем ближе коэффициент к 0, тем меньше связь между признаками, больше погрешность при предсказании значений.

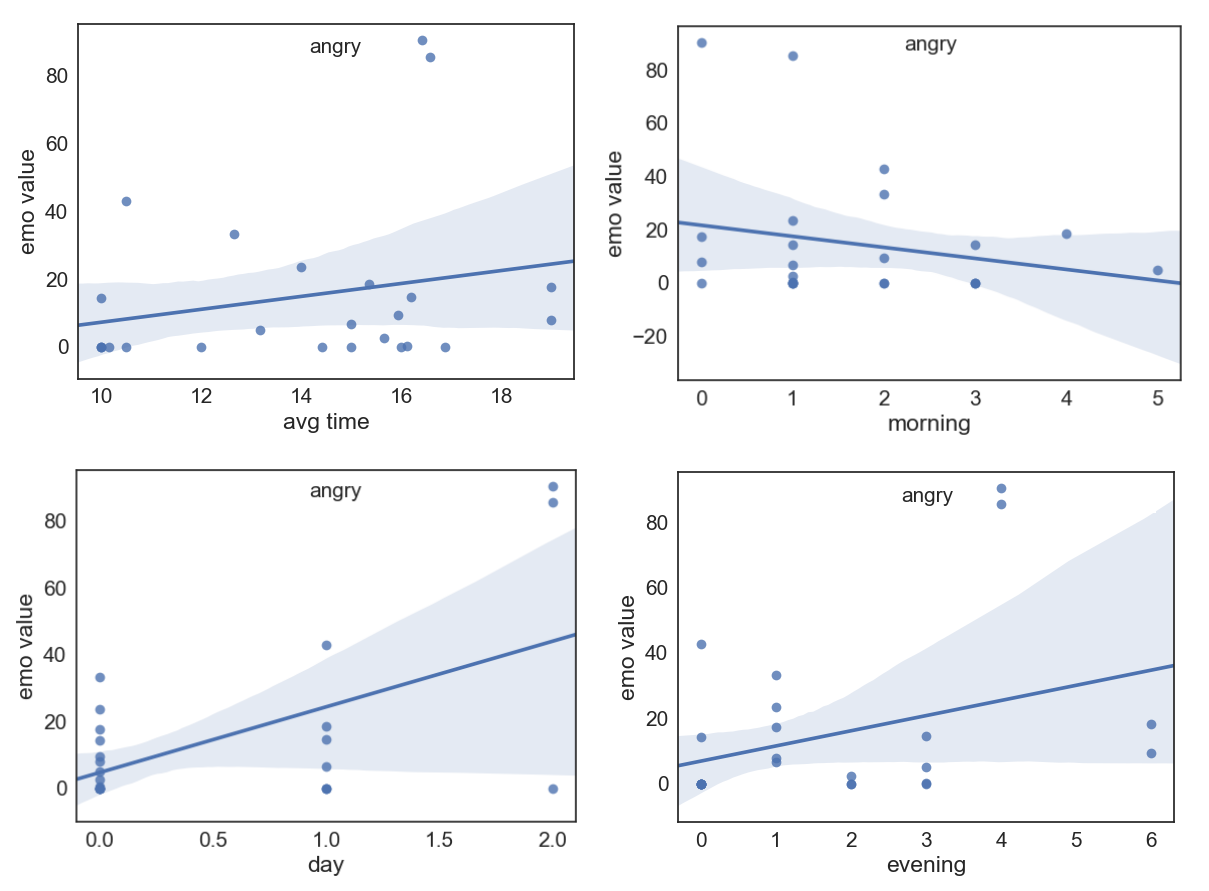
Были выбраны признаки с наибольшими коэффициентами, а значит и сильнее влияющие на эмоции. Также были построены точечные диаграммы для каждой пары признаков и удалены признаки с недостаточным количеством наблюдений, подтверждающих зависимость.

Анализ эмоции «disgust» не проводился, т.к. все ее значения в выборке равны нулю. Отсутствие эмоций (neutral) также не анализировалось и рассчитывалось по формуле   
«100% минус сумма значений по всем эмоциям».

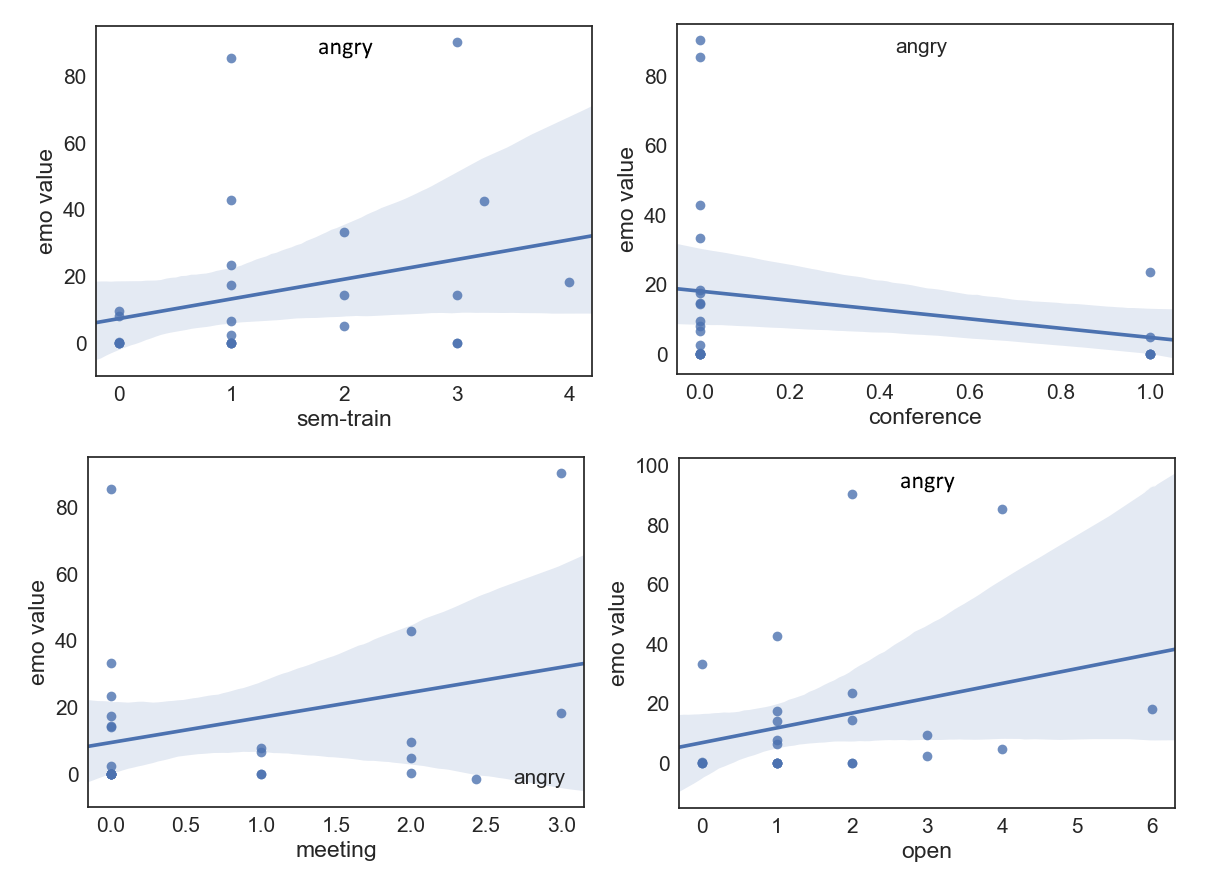
**Злость (angry)**. Признаки: количество дневных и вечерних мероприятий, семинаров, совещаний, открытых мероприятий. Коэффициенты линейной корреляции:



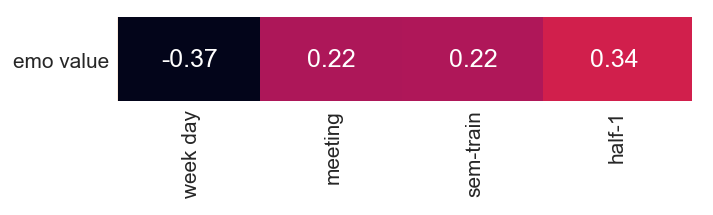
Злость уменьшается с увеличением числа утренних мероприятий, и, наоборот. возрастает с увеличением среднего времени начала, количества дневных и вечерних мероприятий, что можно объяснить человеческой усталостью.



Более сердитые люди, согласно данным, приходят на тренинги, совещания и открытые мероприятия, менее сердитые – на конференции.



24 и 25 апреля уровень злости вырос до 86-90 % (точки хорошо видны на графиках выше), хотя в остальные дни не превышал 45 %. Общее количество наблюдений маленькое, так что эти точки сильно влияют на общую картину, при этом изменяются они независимо от выбранных признаков. При удалении точек влияние времени дня, количества конференций и открытых мероприятий уменьшается. Заметными становятся/остаются следующие признаки:

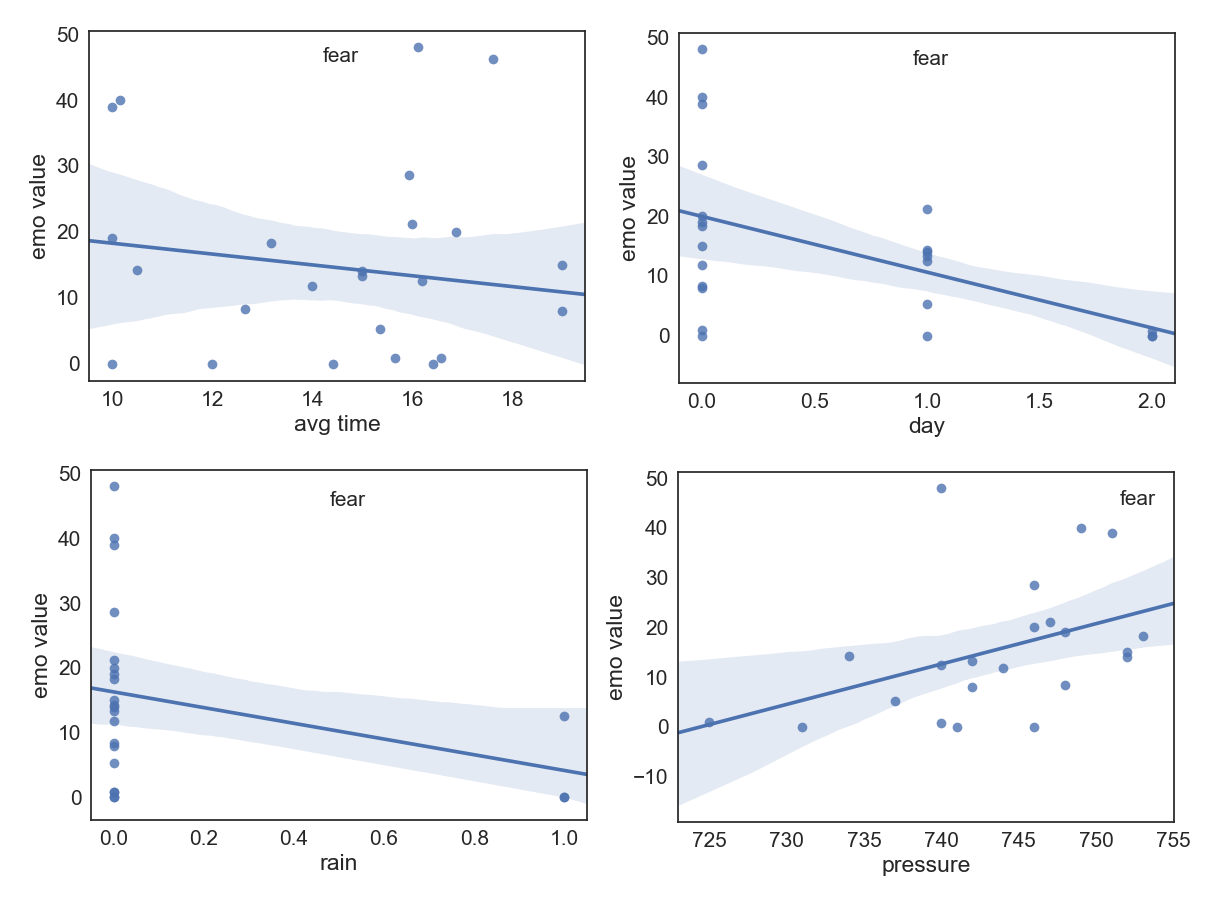


Согласно коэффициентам, злость уменьшается ближе к выходным и больше в первой половине дня (растет при увеличении количества дневных мероприятий), сердиты люди, которые приходят на совещания и тренинги.

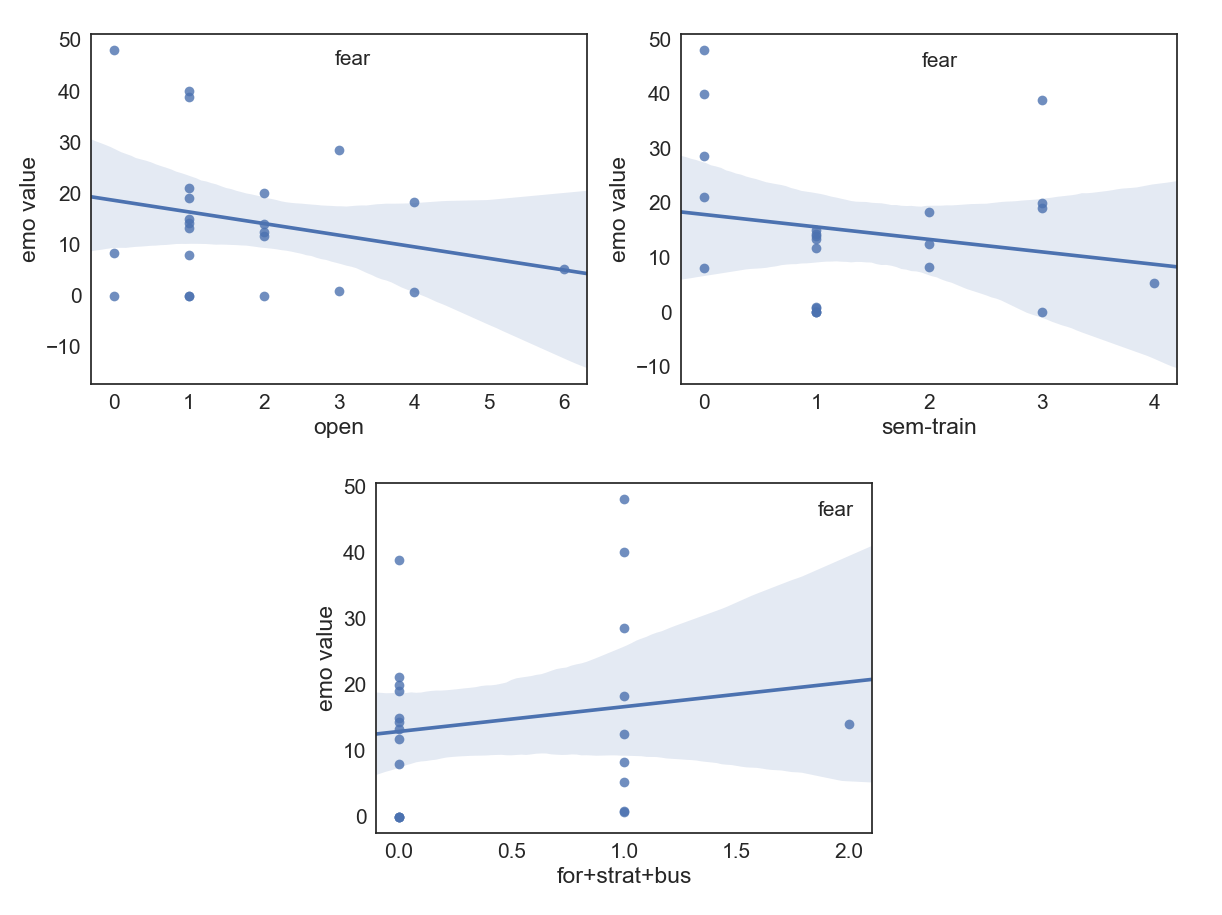
Страх (fear). Признаки: среднее время начала мероприятий, количество дневных и открытых мероприятий, семинаров/тренингов, наличие дождя, атмосферное давление. Признак for+strat+bus – количество (за день) форсайт-сессий, стратегических сессий, бизнес-инкубаторов.



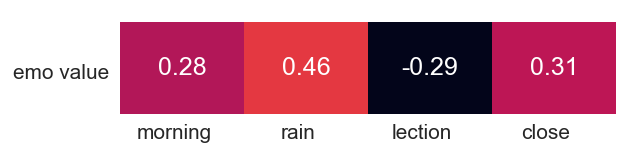
Эмоция, которую программа распознавания определяет как страх, убывает к вечеру, при уменьшении количества дневных мероприятий, при пасмурной дождливой погоде, что может быть следствием сонливости людей, менее открытых глаз и т.п. Страх, напротив, возрастает, с повышением атмосферного давления.



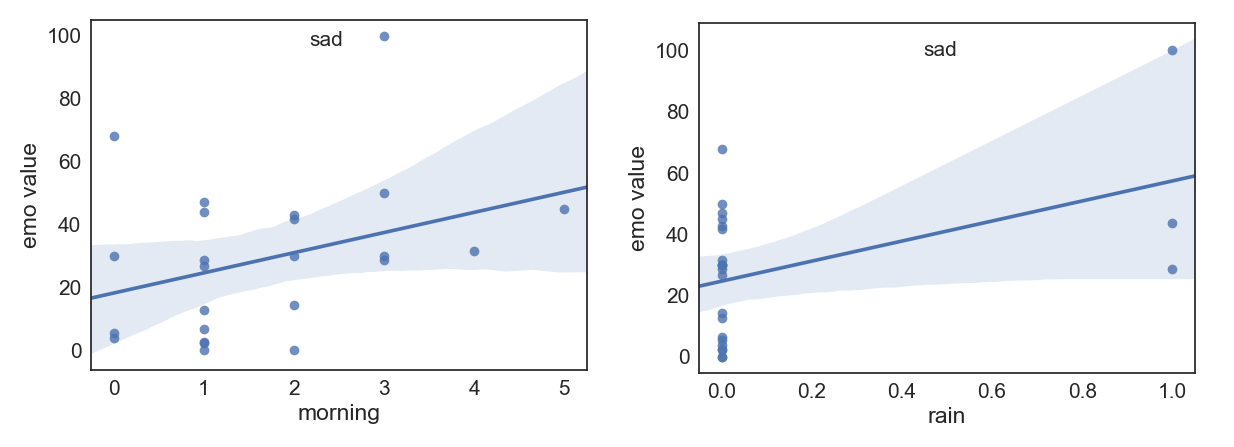
Посетители чувствуют меньше страха, когда приходят на открытые мероприятия, участвуют в семинарах и тренингах, и больше – при участии в форсайт-сессиях, стратегических сессиях, бизнес-акселераторах.



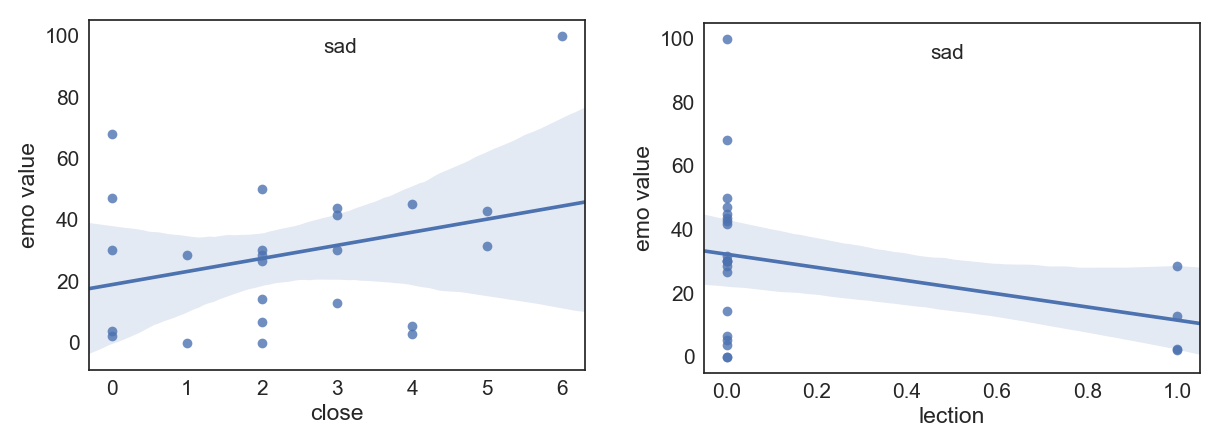
Грусть (sad). Признаки: количество утренних мероприятий, лекций, заседаний рабочих групп, круглых столов, наличие дождя, количество закрытых мероприятий.



Грусть усиливается вместе с количеством утренних мероприятий, и когда на улице пасмурно и идет дождь.



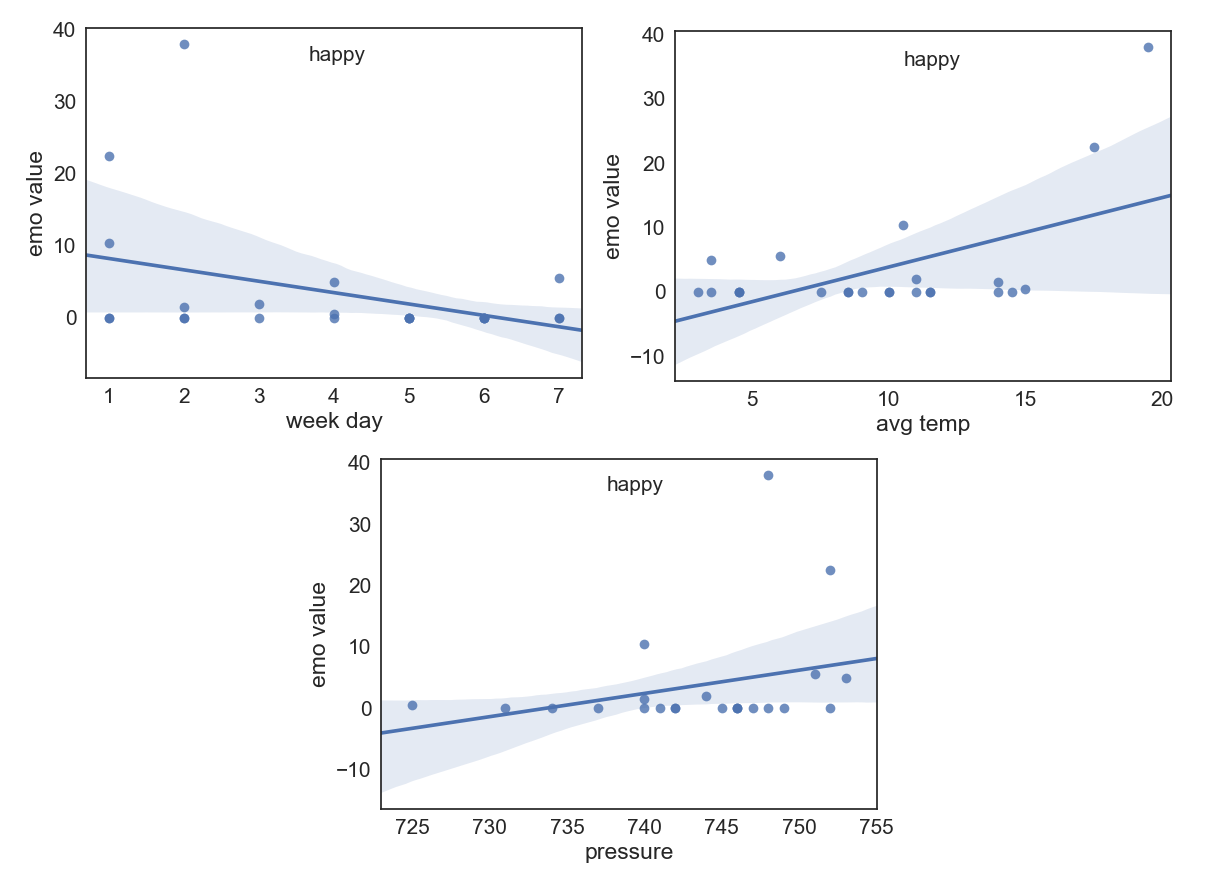
Количество грустных лиц больше при большем количестве закрытых мероприятий за день. Грусть фиксируется реже у посетителей лекций.



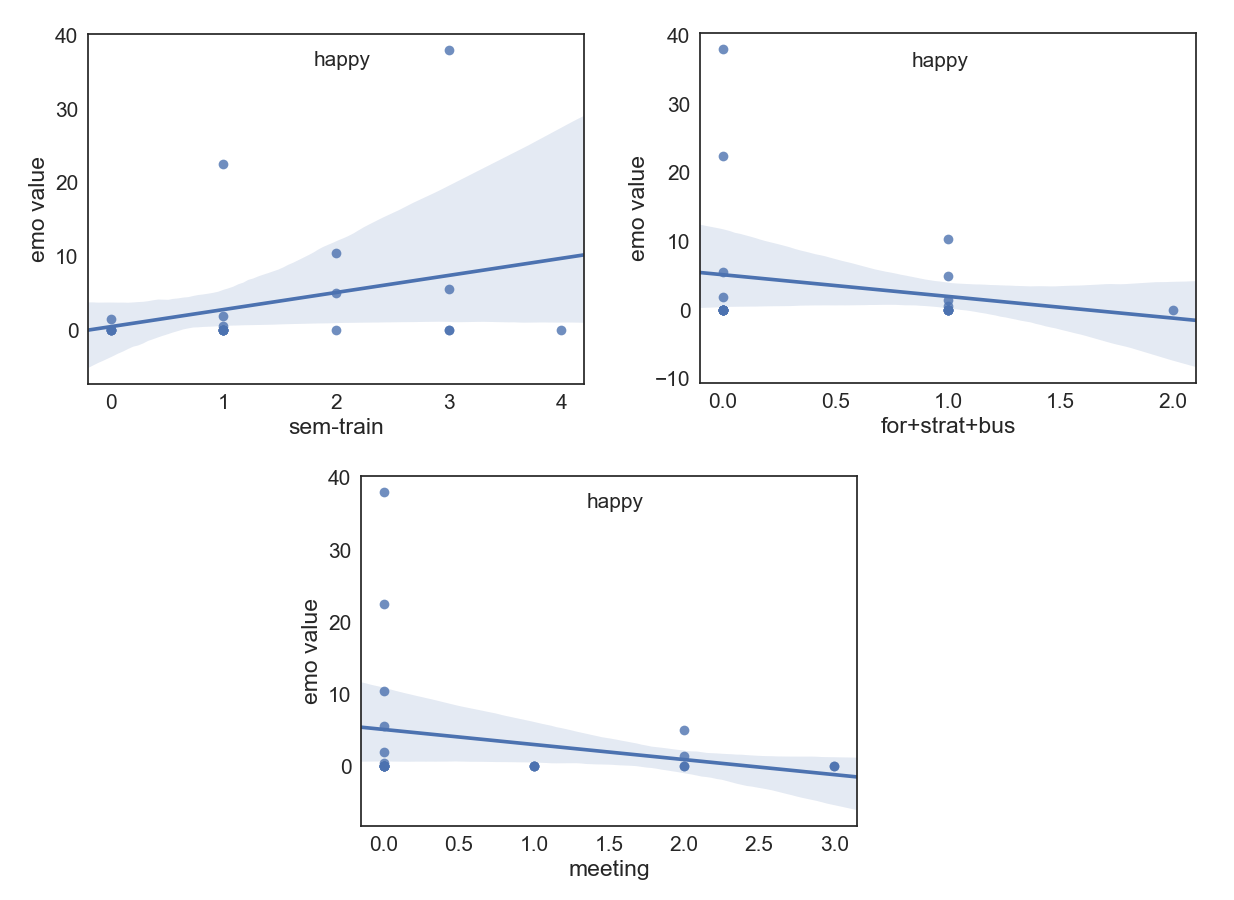
Радость (happy). Признаки: день недели, среднесуточная температура, атмосферное давление, количество совещаний, семинаров/тренингов, форсайт-сессий, стратегических сессий, бизнес-инкубаторов .



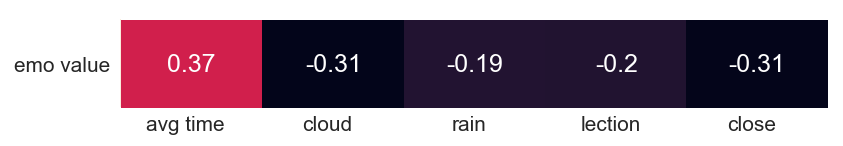
Положительно на количество людей влияют температура воздуха и атмосферное давление. Отрицательно – день недели (ближе к концу недели счастья меньше).



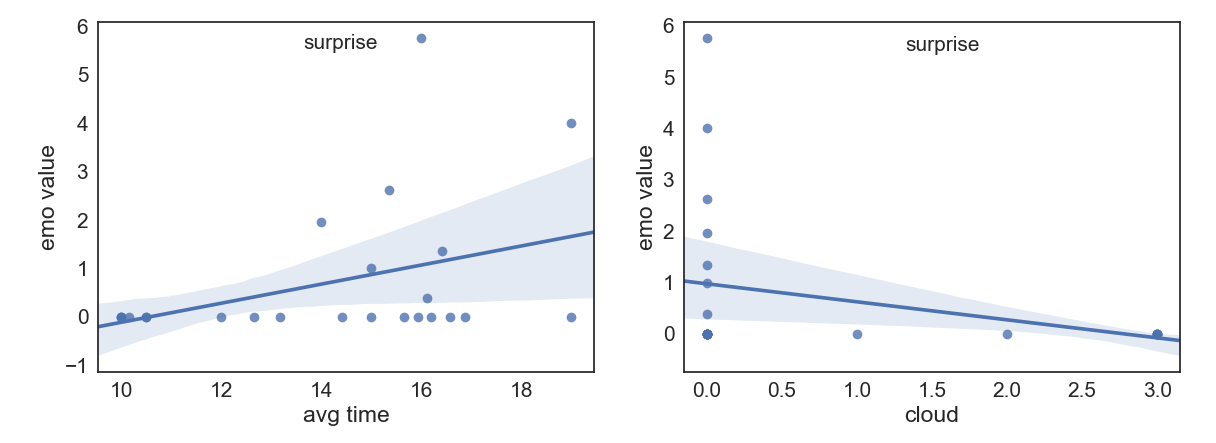
Большее количество счастливых посетителей определяется на тренингах, меньшее – на совещаниях, форсайт-сессиях, стратегических сессиях, бизнес-инкубаторах.



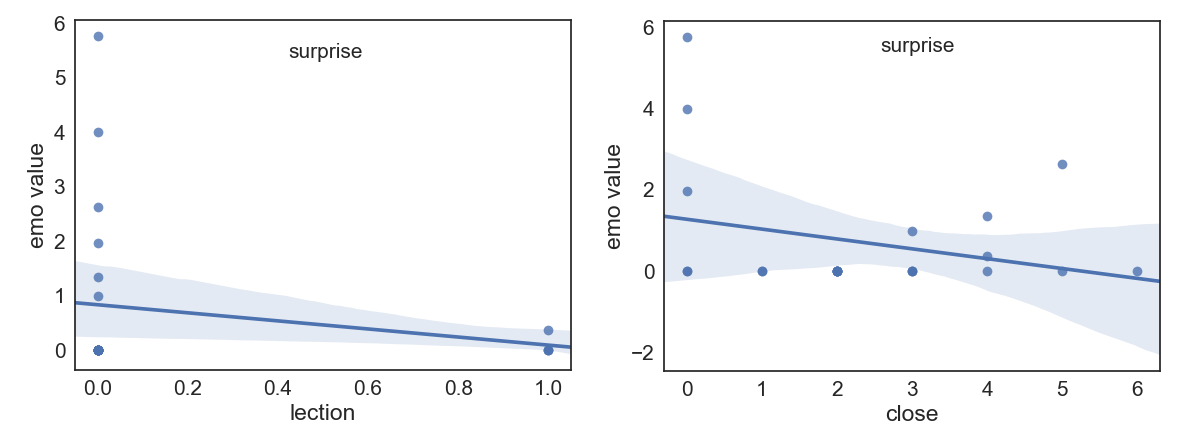
Удивление (surprise). Признаки: среднее время начала мероприятий, облачность, дождь, количество лекций, закрытых мероприятий.



Согласно исходным данным, удивление растет к вечеру и уменьшается   
в пасмурные дни.



Посетители лекций и закрытых мероприятий проявляют мимику удивления реже.



**Построение деревьев решений**

На основе первичного анализа данных были отобраны следующие признаки для прогнозирования эмоций:

|  |  |
| --- | --- |
| Эмоция | Входные признаки |
| angry | week day (день недели), meeting (количество совещаний), sem-train (количество семинаров/тренингов), half-1 (количество мероприятий в первой половине дня) |
| fear | day (количество дневных мероприятий), rain (дождь), pressure (атмосферное давление), sem-train (количество семинаров/тренингов), for+strat+bus (количество форсайт-сессий, стратегических сессий, бизнес-акселераторов), open (количество открытых мероприятий) |
| sad | morning (количество утренних мероприятий), rain (дождь), lection (количество лекций), close (количество закрытых мероприятий) |
| happy | week day (день недели), avg temp (среднесуточная температура), pressure (атмосферное давление), meeting (количество совещаний),  sem-train (количество семинаров/тренингов), for+strat+bus (количество форсайт-сессий, стратегических сессий, бизнес-акселераторов) |
| surprise | avg time (среднее время начала мероприятий), cloud (облачность), lection (количество лекций), close (количество закрытых мероприятий) |

На языке Python были построены решающие деревья (библиотека scikit-learn, DecisionTreeRegressor). Глубина деревьев – 5-8 уровней. Результаты проверялись поочередным исключением каждой даты из обучающей выборки и сравнением разброса (суммы квадратов отклонений).

Прогноз алгоритма по эмоциям на 20 мая: 20 % fear, 40 % sad, 40 % neutral.