**Useful links:**

1. Dataset for the first stage
   1. <https://www.crowdanalytix.com/contests/mckinsey-big-data-hackathon>
2. Datasets descriptions
   1. <https://docs.google.com/spreadsheets/d/1sOQDDGEoChDEx_rLwCB7z_dtRDGXgVE5p3_xS0LJLHI/edit#gid=0>

**Партнёр McKinsey**

Two problems which bothers companies:

1. Что происходит с Биткоином?
2. Где взять data scientist'ов?

4 критерия отбора проектов:

1. Качество алгоритма
2. Точность предсказания
3. Оригинальность идеи

1. Умение донести

**Георгий Мешков, head analytics Gett**

1. Наземный транспорт - 20 триллионов долларов ww
2. Тачками владеют 85% частные лица, 15% такси
3. Низкая утилизация частных машин
4. 850m \* 4% util to 150m at 35% util; less cars - higher average speed
5. 10 min arrival, 20 min trip, 10 min arrial; you pay for the whole trip. Now you pay for all this time. The more compressed are the
6. Gett minimizes time between rides.
7. The question about autonomous cars: who own them - marketplaces OR IT giants OR manufacturers
   1. Answer from the next speaker:
8. Operators have the biggest amount of data about
9. Demand (**это надо минимизировать**) - "число пользователей"
   1. LTV - сколько этот чувак ваще принесёт денег
   2. FTP - first time purchase - когда первый раз воспользовался
   3. RFM - денежный поток от клиента во времени
10. Supply (**это надо максимизировать**) "таксисты, 'предложение' клиентов"; number of empty cars we have
    1. GH (getting hour) - к разговору о полезном/бесполезном времени
    2. Utilization - сколько поездок в час
    3. Acceptance rate - таксисты либо пьют чай
    4. Completion rate - сколько заказов вы обеспечили машинами?

**Евгений Бельдорф**

1. Three parts of every journey:
   1. We suggest him a route to drive during waiting to maximize the probability of getting the next request ASAP "dynamic fleet management"
   2. "Dynamic fleet management" philosophy
      1. AI based
      2. Increased utilization
      3. AV ready (this technology is super important for AVs since drivers know what to do and cars don’t)
      4. Piot results:
         1. +15% income
         2. +22% utilization

**Next speaker, chief data scietist in Gett (see him on the site)**

1. For AVs the reward is "passenger doesn’t die'
2. The problem you have to thing about too many things in advance
3. Gett search
4. "Data driven utilization optimization"
5. In Gett they have states for Moscow (Moscow is split into regions)
6. Reward for Gett "net search completion ratio". Gett uses reinforcement learning to increase the net completion rate.

**Inputs**

**7 файлов**

driver paths - про ДРУГОЙ город

1. Как водилы едут (Москва + КАКОЙ-ТО город)
2. Заказы (Москва)
3. Что пользователи вбивают в поиск
4. Отзывы пользователей о поездках

**Tasks**

1. **ANY task with the data**
2. Картографические сервисы
3. Предсказание заработка
4. Классификация адресов
5. Тэгирование отзывов
6. Предсказание LTV
7. Dynamic fleet management

**Завтра в 10:30 презентуем решения, рисовать до этого момента не нужно**

Вопросы по данным:

1. Датасет 1
   1. Чем кроме города и отсутствия метрики скорости отличаются датасеты 1.1 и 1.2?
2. Датасет 2
   1. orders\_drivers / orders\_riders - таблицы друг с другом не бьются (у клиентов и водителей своя жизнь)
3. Датасет 3
   1. Отзывы, просто отзывы (на разных языках)
   2. Как вы сейчас с ними работаете?
4. Датасет 4
   1. Откуда и куда заказывают такси на протяжении недели с **часовой** точностью

**Экономический эффект от supply prediction'а**

**Passive**

Водители рассказывают, что при выходе приложение часто напоминает им, что они остановились в шаге от определенной суммы (Каждый раз после нажатия на кнопку «Выйти из приложения», система присылает уведомление из серии «Вам осталось всего $10 до $330. Вы действительно хотите выйти?».) или от суммы заработка по сравнению с тем же днем на прошлой неделе»,

«Например, у Uber есть алгоритм … В случае с Uber водители получают следующий заказ еще до того, как завершен предыдущий».

Consider an algorithm called **forward dispatch** — Lyft has a similar one — that dispatches a new ride to a driver before the current one ends. Forward dispatch shortens waiting times for passengers, who may no longer have to wait for a driver 10 minutes away when a second driver is dropping off a passenger two minutes away.

Nikita Tolstoy, [09.12.17 22:06]

<https://vc.ru/22971-mental-uber>

Nikita Tolstoy, [09.12.17 22:33]

(о том, как работает механизм Убера) <http://taxiuber.ru/forum/threads/675/page-2>

"короче, сделал я эти поездки, но меня мало-мало "кинули". не выплатили ничего, на запрос ответили: "ну вы приняли же меньше 80% заказов"...

да, я принял 70%.

но я теперь разборчив.. кому надо иметь проблемы с клиентом, рейтинг которого ниже 4.6? или: приходит заказ, я стою в пробке 5 рядов на Б.Каменном, и надо разворачиваться и катить обратно?? или - надо ехать к клиенту из гарантированной зоны в негарантированную, да ещё больше 10 минут в пути? как говорил Буратино: "ищи дурака!"

считаю этот барьер (80%) искусственным и дискриминационным, во."

At any moment, the app shows drivers how many trips they have taken in the current week, how much money they have made, how much time they have spent logged on and what their overall rating from passengers is. All of these metrics can stimulate the competitive juices that drive compulsive game-playing.

<https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html?_r=0>

**5 Devious Tricks Uber Uses to Keep Drivers on the Road**

5th: Not Disclosing a Passenger's Final Destination Until the Ride Is Accepted

<http://www.complex.com/life/2017/04/uber-5-tricks-keep-drivers-on-road>

По статистике, в режиме «Домой» заказов приходит меньше обычного. **Но заказы есть, и они очевидно выгодные.** Опыт нескольких тысяч водителей, что каждый день пользуются опцией «Домой», — тому подтверждение.

<https://driver.yandex/faq-order/>

**Have a "I'm done after this delivery" button so you don't have to reject jobs**

<https://www.coworker.org/petitions/have-a-i-m-done-after-this-delivery-button-so-you-don-t-have-to-reject-jobs>

Basically, if we can predict the amount of time a merchant will spend preparing the food, we can offer the delivery to a driver who is currently working on another delivery, such that when they finish their current dropoff, the food will be ready for them at the next pickup location.

<https://blog.postmates.com/making-postmates-even-more-efficient-690aacb7fad6>

**Ver 1.**

1. We found a hole in your pocket: some customers never get a taxi (all the people around just reject it).
2. What if there was a car nearby? Well… it may be there if we think an advance.
3. Let us present the algorithm to predict the driver leaving work in advance, predicting the money loss and acting proactively. Uber uses some mind tricks like "do you really want to leave us?", but as soon as everyone knows the joke nobody will ever laugh it. The competitive advantage of this technology is that
4. Two approaches:
   1. We somehow make driver to spend more time at work
   2. We negotiate the conditions with passenger like "are you ready to
   3. We compensate the missing driver with a vacant one in advance

**Ver 2.**

1. Supply prediction
   1. In very generic form, the problem Gett loses customers - supply not meeting the demand at some points of time. Based on the data we received, the cost of such a miss can be X$ a month for Moscow only .
   2. Predicting the supply is a too rocky mountain to climb with such a contracted dataset, today we make just a step towards there, predicting the decrease in supply by **identifying the drivers who are going to leave.**
   3. <about technology>
   4. This thing is better than Uber's "do you really wanna leave" stuff because of the fundamental problem of all psycho tricks - they work until you don't know you get fooled. Instead, Gett can utilize more proactive approaches:
      1. Demand is geo-dispersed thing: we can ask the car from nearby to get closer to the area with "going to decrease" demand in advance
      2. "Wise" simulation by giving a driver a premium for staying at work (premium is limited by the maximum saved revenue loss"
   5. Next steps: take data to start building full monitoring algorythm

**Ver 3.**

1. Supply prediction
   1. Supply doesn't meet a demand. Risk = X$
   2. Predicting supply - we lack data (not consistent data about drivers).
   3. Predicting decrease in supply - not a problem.
   4. About technology
   5. Uber sucks since it's passive system.
   6. If we build active system
      1. Prevent the "demand" issue
      2. "Wisely" keep a customer on track
   7. Next steps: more data

**Ver 4.**

1. Supply prediction
   1. Supply <> demand. Risk = 3% revenue.
   2. Big dream: predicting supply.
   3. Predicting DECREASE in supply - ok with this data.
   4. Technology
   5. Uber - passive
   6. Gett - active:
      1. Prevent the issue
      2. Wisely" keep the driver
   7. We need more data

a. In very generic form, the problem Gett loses customers - supply not meeting the demand at some points of time. Based on the data we received, the cost of such a miss can be X$ a month for Moscow only .

b. Predicting the supply is a too rocky mountain to climb with such a contracted dataset, today we make just a step towards there, predicting the decrease in supply by identifying the drivers who are going to leave.

c. <about technology>

d. This thing is better than Uber's "do you really wanna leave" stuff because of the fundamental problem of all psycho tricks - they work until you don't know you get fooled. Instead, Gett can utilize more proactive approaches:

i. Demand is geo-dispersed thing: we can ask the car from nearby to get closer to the area with "going to decrease" demand in advance

ii. "Wise" simulation by giving a driver a premium for staying at work (premium is limited by the maximum saved revenue loss"

Next steps: take data to start building full monitoring algorythm

**Данные о Gett в России**

Ранее основатель и руководитель Gett Дэйв Вайсер заявлял, что на сегодня доля российского рынка в общем бизнесе Gett составляет около 25%.

**Supply prediction - Feedback on our solution from Ianir Yanirs**

1. Technology was one of the best; the level of what you showed was very high.

2. Business problem statement was weak.

3. Presentation was weak.

Driver stimulation does not work much; Uber experimented with rewarding, ended up with the decrease in trips quality. Better you find the trip on his way back home.

When we described other possible applications of the supply prediction technique - he said that would work "but you have to pitch it".

**Final presentations**

1. **18-kenna-… 0**
   1. Gett uses shadow customers, why not to substitute with model?
   2. Fraud/not fraud is binary classification problem.
   3. 7 days of driver + 2 days of another driver in mix. LightGBM <https://github.com/Microsoft/LightGBM>
   4. **Generally speaking, model tracks changes in behavior**
   5. Model won't work on people who are constantly driving fast.
2. **Kanape +**
   1. ШАД guy
   2. Анимация на слайдах :)
   3. Две идеи:
      1. Давайте поменяем распределение водителей так, чтобы оно лучше мапилось на распределение заказов
      2. Давайте сделаем Телеграм-бота для приёма отзывов.
   4. "Idea is to reposition the supply?" - "Yes"
   5. "What is the cost function?"
   6. "Any thoughts on how to make drivers to use your commands?"
   7. "What is the economical value of implementing your solution?"
   8. "On reviews classification: what are the insights?"
3. **Accuracy is not ROC 0**
   1. Decide which drivers have now taken an order from another marketplace.
   2. "What is the solution for the problem "a driver leaked to another marketplace" - don't know
   3. "Can you please show the numbers for the model?"
4. **Supersonic Potato -1**
   1. Driver gamification (like Uber does)
   2. Markov chain to predict driver's revenue
5. **МакдакHack ++**
   1. ML-driven geo-recomendation system
   2. Classifying places by "complex to get" and "easy to get"
   3. No model
   4. *Model: k-means clusterization and decision rules algorithm are used to classify complicated and simple locations; Complicated locations are highlighted on the map, colored by type*
   5. "Why is it better?"
      1. Let's suggest complex strategies for users; Uber and Yandex
   6. "How far is your solution from your vision?"
   7. 4/5 членов команды - сотрудники Double Data. Ребята принесли решение, основывающееся на кроулинге кучи данных с публичных источников.
6. **BuzzwordLearning + по модулю английского**
   1. Слишком широко начал. Можно быть более конкретным.
   2. Модель: e-greedy algorythm with Cat Boost exploitation formula
   3. Предсказание направления движения
   4. "Thank us all"
7. **EORA ---**
   1. Market analytics based on numbers in the dataset
   2. Фича: одновременный заказ многих такси
   3. "Какие конкретно юзкейсы?"
   4. "How did you use the data?" … "So how did you use the data?"
8. **Getto 0** 
   1. Dataset 4.
   2. Similar to the team Kanape: давайте оптимизируем supply под demand.
9. **Noobie Students +**
   1. "Smart carpooling": predictict groups of people going to "pretty the same" area
   2. "We used OpenCV..." -> "Why did you need computer vision?"
   3. **BUT!** they solved the problem "find the city"
10. **SK +**
    1. Super clear problem statement, super clear model expanation, super on-time
    2. Подача "ровная"
11. **Igor & Co +** 
    1. Clusterization + adding data on weather (input as just one of features) + teaching the model with xgboost + k-folds learning
    2. "What algorythms are also good except yours?"
    3. Очень много хороших глубоких вопросов. Но команда более сбалинсированная.
12. **Welcome to Uganda +**
    1. Most positive presentation.
    2. Here is where machine learning should shine a lot
    3. Crawled the reviews in web (Google Play market)
    4. Sentiment analysis on the increased amount of reviews.
    5. Чуваки зарешали категоризацию ваще на раз.