# Описание задачи и набора данных

Было решено выбрать в качестве задачи прогнозирование затрачиваемых средств на различные виды услуг.  
В связи с конфиденциальностью данных о банковских транзакциях было принято решение о написании алгоритма генерации данных.  
Банковская транзакция – это Id отправителя, id получателя, id транзакции, время совершения транзакции, сумма транзакции.  
Для генерации так же использовались данные о количестве денежных средств в начале месяца, количестве денежных средств перечисленных на данный счёт (будем считать это зарплатой).   
Количество денежных средств к началу месяца и количество денежных средств перечисленных на счёт образуют капитал, который клиент банка использует для банковских транзакций. Значит эти денежные средства прямым образом влияют на транзакции производимые клиентом.   
В итоге алгоритм должен сгенерировать двумерный лист  
 [user\_id; purchase\_type; transaction\_id; sum; month]  
Вместо id получателя тип покупки, это не случайно, так как выявление какую услугу предоставляет получатель денежных средств не задача данной работы, поэтому для простоты мы просто будем генерировать тип услуги.  
Вместо времени совершения транзакции указано слово «месяц», для данной задачи не имеет значения в какое время и в какой день была совершена транзакция. А парсинг месяца из даты типа (дд.мм.гггг; чч:мм) задача тривиальная.

# Преобразование данных

Мы имеем двумерный лист, строкой которого является коллекция типа :  
[user\_id; purchase\_type; transaction\_id; sum; month]  
Для прогнозирования затрачиваемых средств на различные виды услуг в следующих месяцах нам не нужно знать id транзакции и id клиента банка. Но нам нужно знать тип услуги, количество денежных средств потраченных на данную услугу и месяц, в который данные средства были потрачены.  
Так же нам понадобится информация о количестве денежных средств, которыми распоряжался клиент банка. Данную информацию соберём во время генерации. При наличии настоящих данных нужно будет пробегать по всем транзакциям и считать, но в нашем случае мы можем сделать это во время генерации.  
В итоге преобразование данных сводится к изъятию из созданного генератором набора данных информации из трёх столбцов, сложения сумм транзакций конкретного типа услуги и месяца, а так же конкатенации с информацией о количестве имеющихся денежных средств.  
В итоге у нас получился двумерный список. С помощью библиотеки NumPy преобразуем список в матрицу и с помощью метода hsplit() разобьём матрицу на две, по которым и будем обучать машину.

# Машинное обучение

Данная задача относится к задаче прогнозирования с учителем. Мы будем по вектору пытаться предсказать количество затраченных средств на определённый тип услуг

Машинное обучение будет проводиться по алгоритму «случаеного леса» (Random Forest).

«Случайный Лес» — алгоритм, который для полученных данных случайным образом создает множество деревьев принятия решений и потом усредняет результаты их предсказаний. Алгоритм построения дерева очень быстр, поэтому не составляет большого труда сделать столько деревьев, сколько будет нужно. Он почти не требует конфигурации, что упрощает его использование на начальном уровне. При этом он дает довольно точные результаты.

# Тесты

Из-за того что данные, по которым обучалась машина, созданы искусственно точность прогнозирования узнать достаточно тяжело.

# Выводы

Выполнение данного проекта позволило углубиться в понимание машинного обучения. В ходе работы над поставленной задачей был создан алгоритм прогнозирования затрачиваемых средств на различные виды услуг. Однако для определения насколько он точен нужно провести еще кое-какие работы, а именно проанализировать данные о транзакциях настоящих людей, желательно чтобы их заработок был различный и имелись представители разных полов и возрастов.  
Помимо проблемы с данными перед выполнением поставленной задачи встала проблема классификации предоставляемых услуг. Для решения этой проблемы нужно сопоставлять id пользователя и тип услуг, которые он предоставляет.