EPFL
EXTENSION
SCHOOL

Online grocery shopping

Capstone project for EPFL Extension School Applied Data Science – Machine Learning

Author: David da Paz

Date: January 14, 2019

EPFL EXTENSION SCHOOL

instacart

Outline

- Instacart, a brief introduction
- Business Model
- Revenue Model
- Presentation of the problem
- Business implications
- Dataset
- Toolbox
- Exploratory and Descriptive Analysis
- Machine Learning, a first approach
- Features Engineering
- Machine Learning, second approach
- Conclusion
- Appendix

EPFL EXTENSION SCHOOL

instacart

Instacart, a brief introduction

- **Instacart** is an American company that operates as a sometimes-same-day grocery delivery service. Customers select groceries through a web application from various retailers and the order is delivered by a personal shopper.
- Instacart's service is mainly provided through a smartphone app, available on iOS and Android platforms, apart from its website. Some of its participating stores allow access through their own websites on a desktop browser.

EPFL EXTENSION SCHOOL

instacart

Business model

1. A customer places his order for groceries and pays online to instacart.
2. A personal shopper receives the order and starts collecting items as mentioned in the order.
3. Shopper pays the bill through Instacart's prepaid debit card which is accepted at the store
4. The shopper then goes to deliver the groceries to the customer as per the address mentioned in the order.

Any tip paid in cash at the time of delivery directly goes to the shopper and the tip paid to shopper during checkout gets accumulated in his Instacart account and is paid at the end of the week along with the salary

EPFL EXTENSION SCHOOL

instacart

Revenue model

- Delivery fee
  - Every order processed by Instacart which is above the value of $35 attracts a standard delivery fee of $3.99 for a scheduled or 2 hour delivery and $5.99 for a 1 hour delivery
  - Orders under $35 value are charged at $7.99 for a scheduled or 2 hour delivery and $9.99 for a 1 hour delivery.

- Membership fee (Instacart Express)
  - Instacart offers an annual membership by the name ‘Instacart Express’ priced at $149 (up from $99). Users having this membership can get free delivery of groceries for full 1 year with few terms and conditions.

- Mark up prices
  - Some stores selling their products on Instacart offer the same prices as their in-store prices but few stores listed on Instacart have a mark-up of 15%+more from their in-store prices. The revenue from these mark-up prices goes to Instacart which helps them pay the shoppers.

EPFL EXTENSION SCHOOL

instacart

Presentation of the problem

- Using a online grocery shopping dataset, the idea is to find interesting patterns in the data
- Build a model to predict which grocery items each Instacart user will likely reorder based on the user's purchase history
- It is a binary classification problem
- The target variable: a binary variable for whether or not the user reordered the product

EPFL EXTENSION SCHOOL

instacart

Business implications

- With the proposed predictive model, Instacart could provide its users with intelligently targeted purchase recommendations, discounts, or other promotions
- This could help Instacart:
  1. Improve customer user experience
  2. Help retaining customers
  3. Increase the number of purchases → Increase revenues

EPFL EXTENSION SCHOOL

instacart

Dataset

- ”The Instacart Online Grocery Shopping Dataset 2017”
- Anonymized dataset contains a sample of
  - over 3.4 million grocery orders from more than 200,000 Instacart users
- Concerning the 3.4 million grocery orders
  - About 3.2 million are orders prior to that users most recent order, and
  - about 131 thousand are users’ most recent orders
- Six relational tables
  - Aisles, departments, order_products_prior, order_products_train, orders and products

EPFL EXTENSION SCHOOL

instacart

Toolbox

Getting the data → Data Preparation → Exploratory & Data Analysis → Features Engineering & Machine Learning

EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis

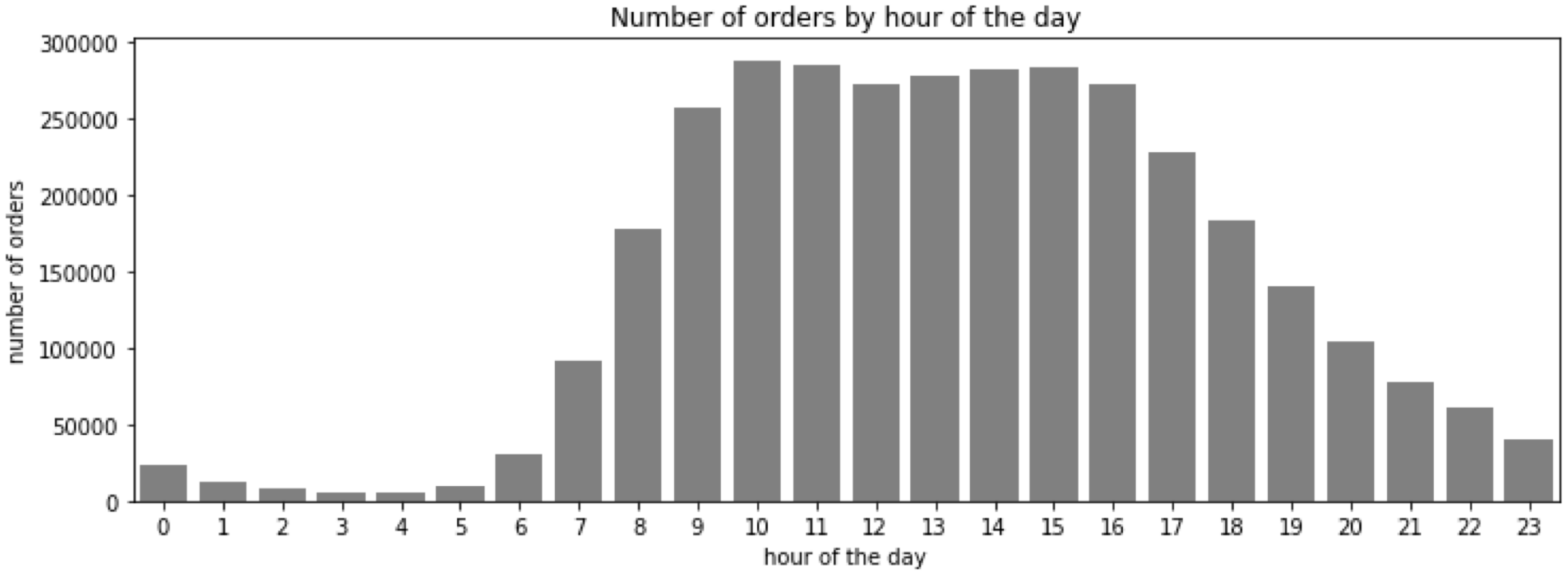
Day 0 and day 1 are the days with most orders within the week

Although nothing is mentioned in the dataset, day 0 and day 1 might correspond to the weekend days, Saturday and Sunday

EPFL EXTENSION SCHOOL

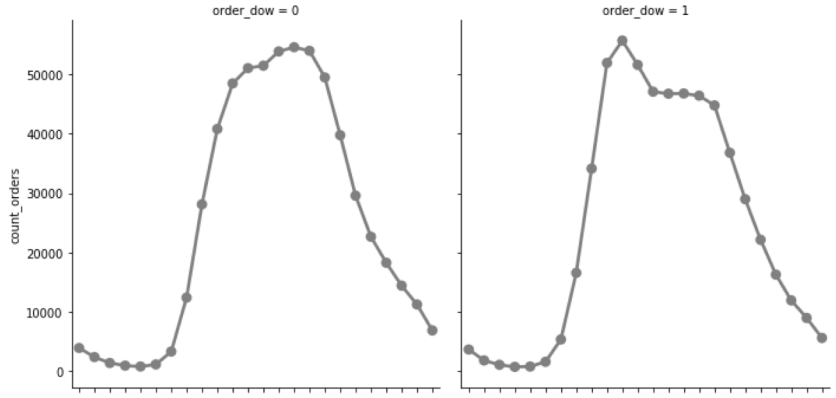
instacart

Exploratory and Descriptive Analysis

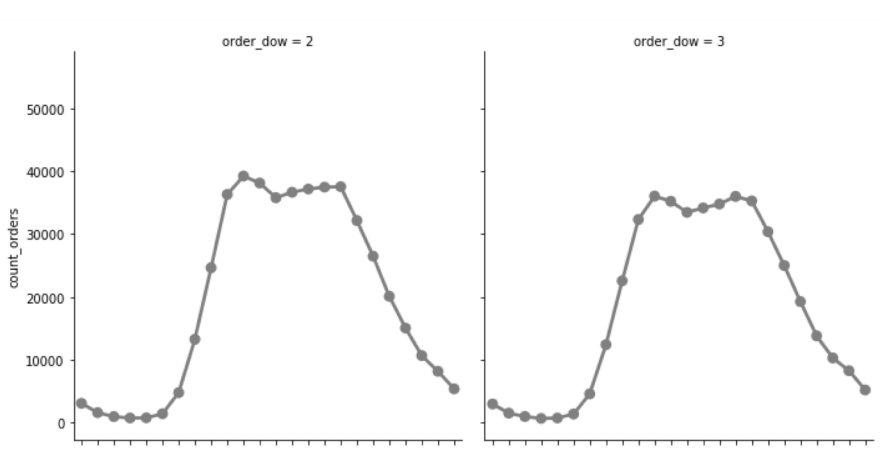


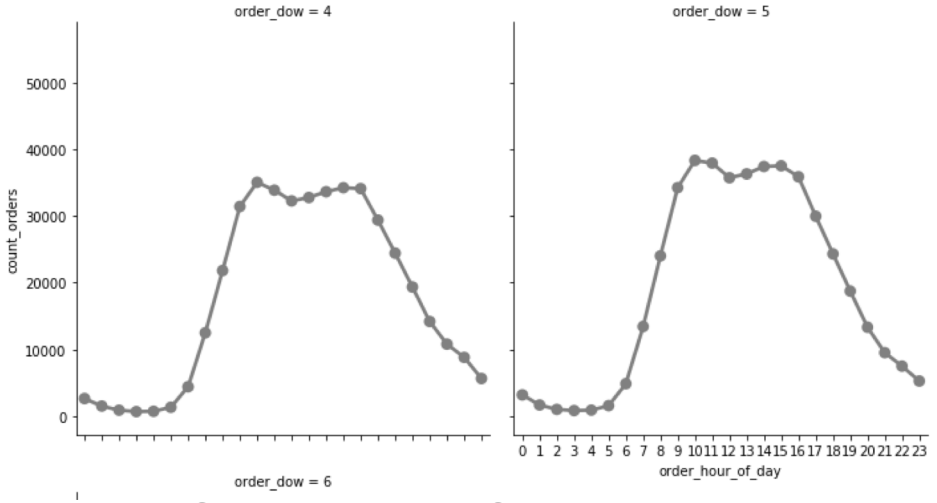
Concerning the hours of the day, most of the orders are made during working hours (from 8:00 to 19:00)

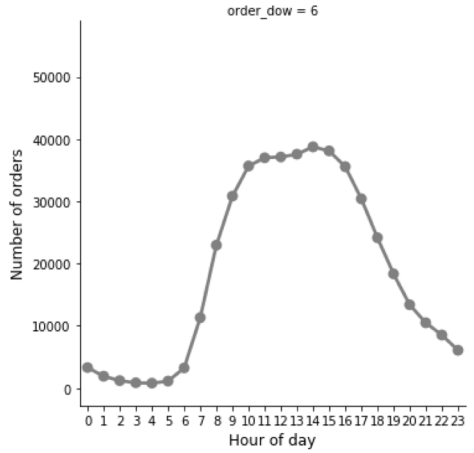
Exploratory and Descriptive Analysis



The pattern of hours is different for day o and day 1 versus day 2 to day 6



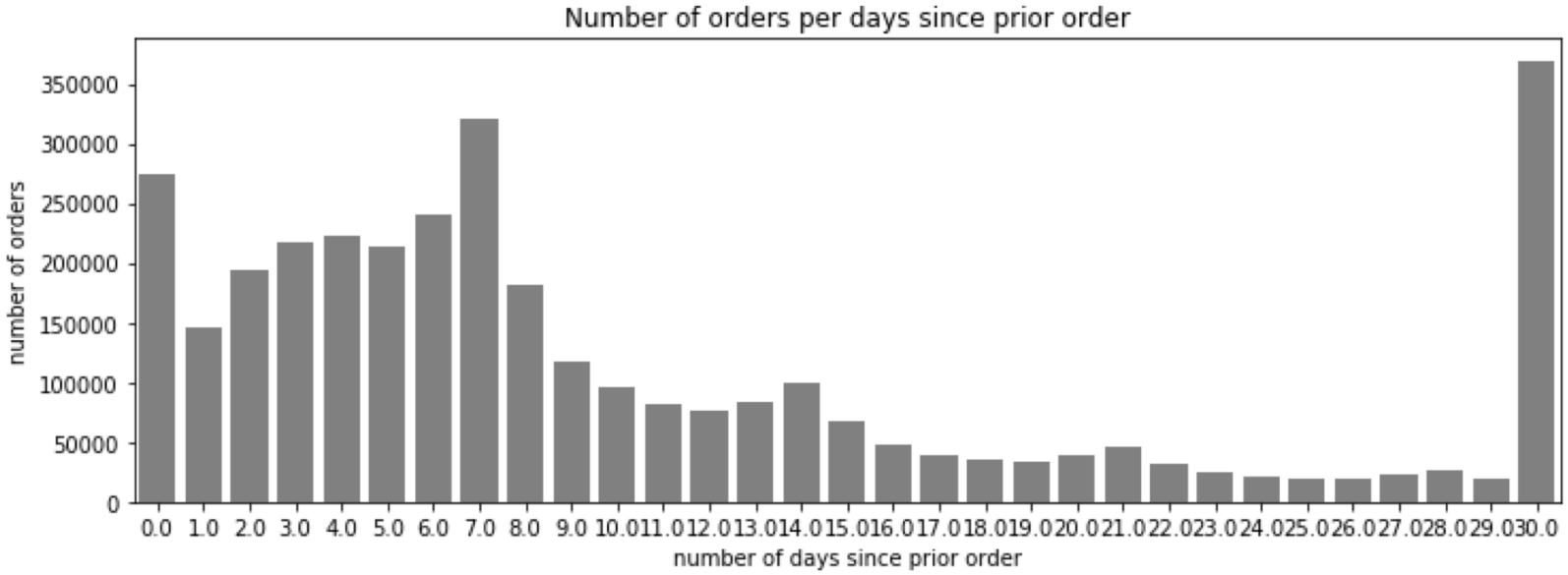




EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis



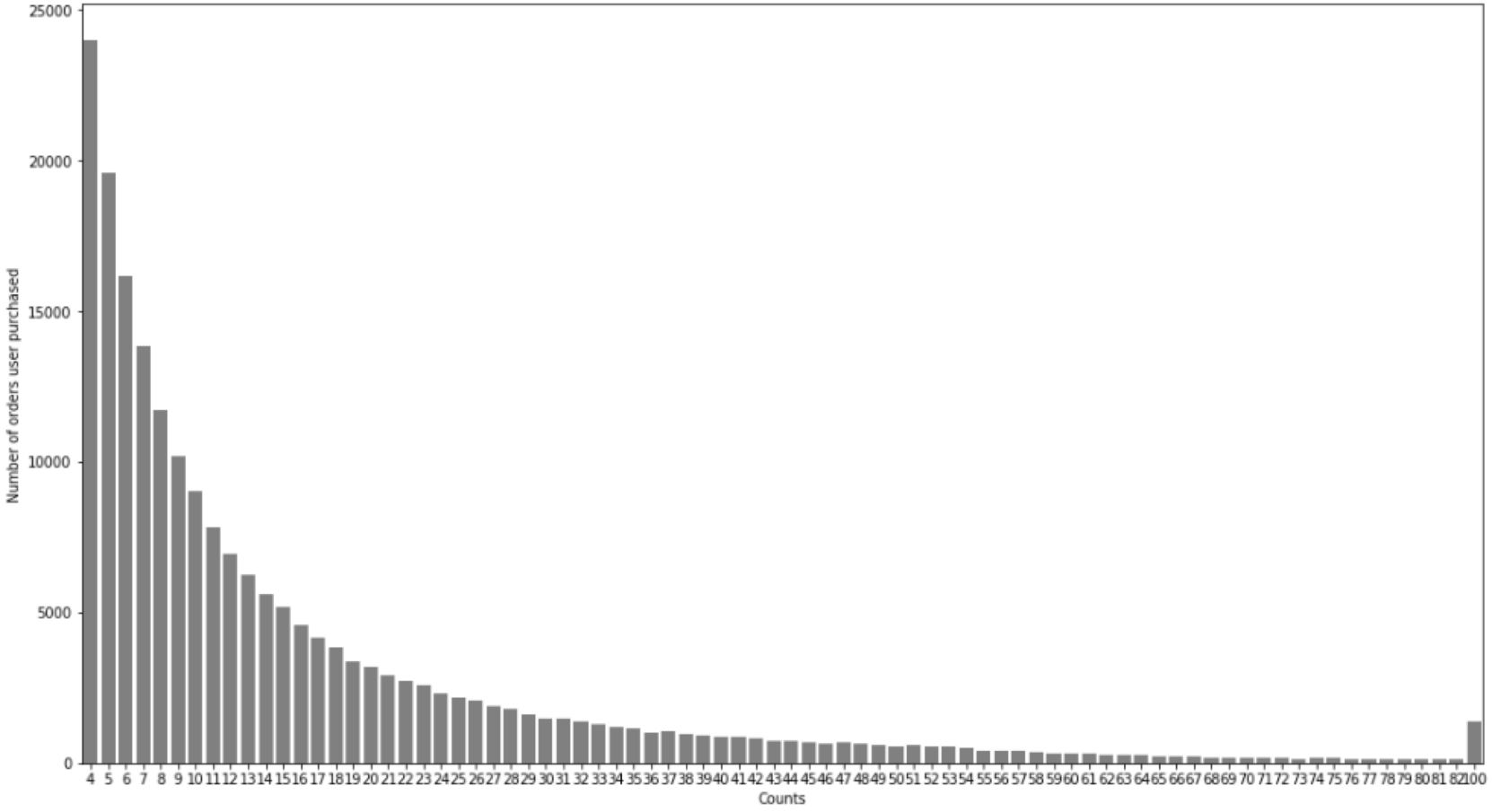
Most of the orders are made at least 7 days since the prior order

The 30 days since prior order might be related with monthly orders from households, businesses, offices

EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis



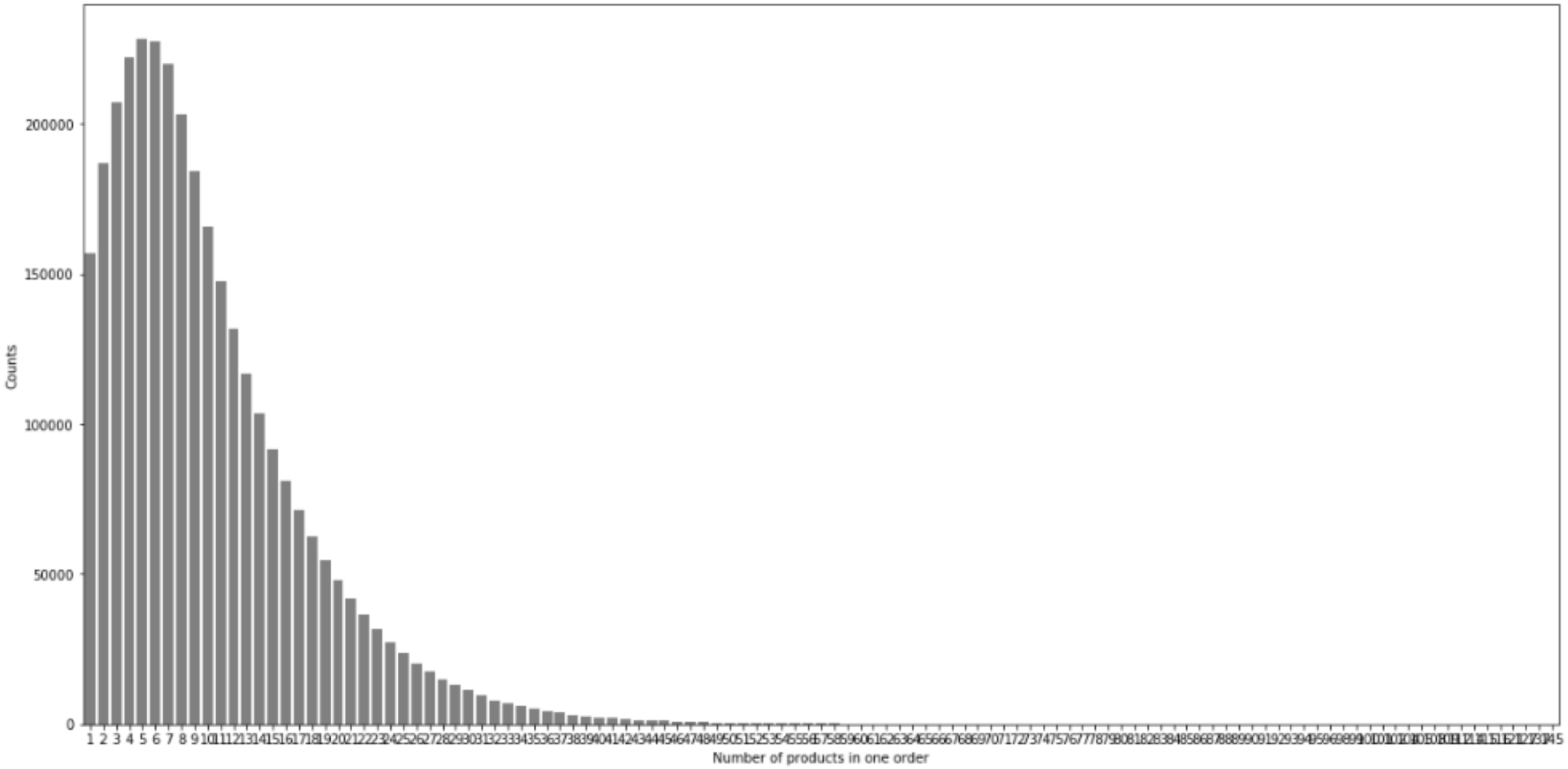
4 orders is the minimum order made by a user and this is the more common

From 4 orders we see a decrease until we have some users that ordered 100 times

EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis

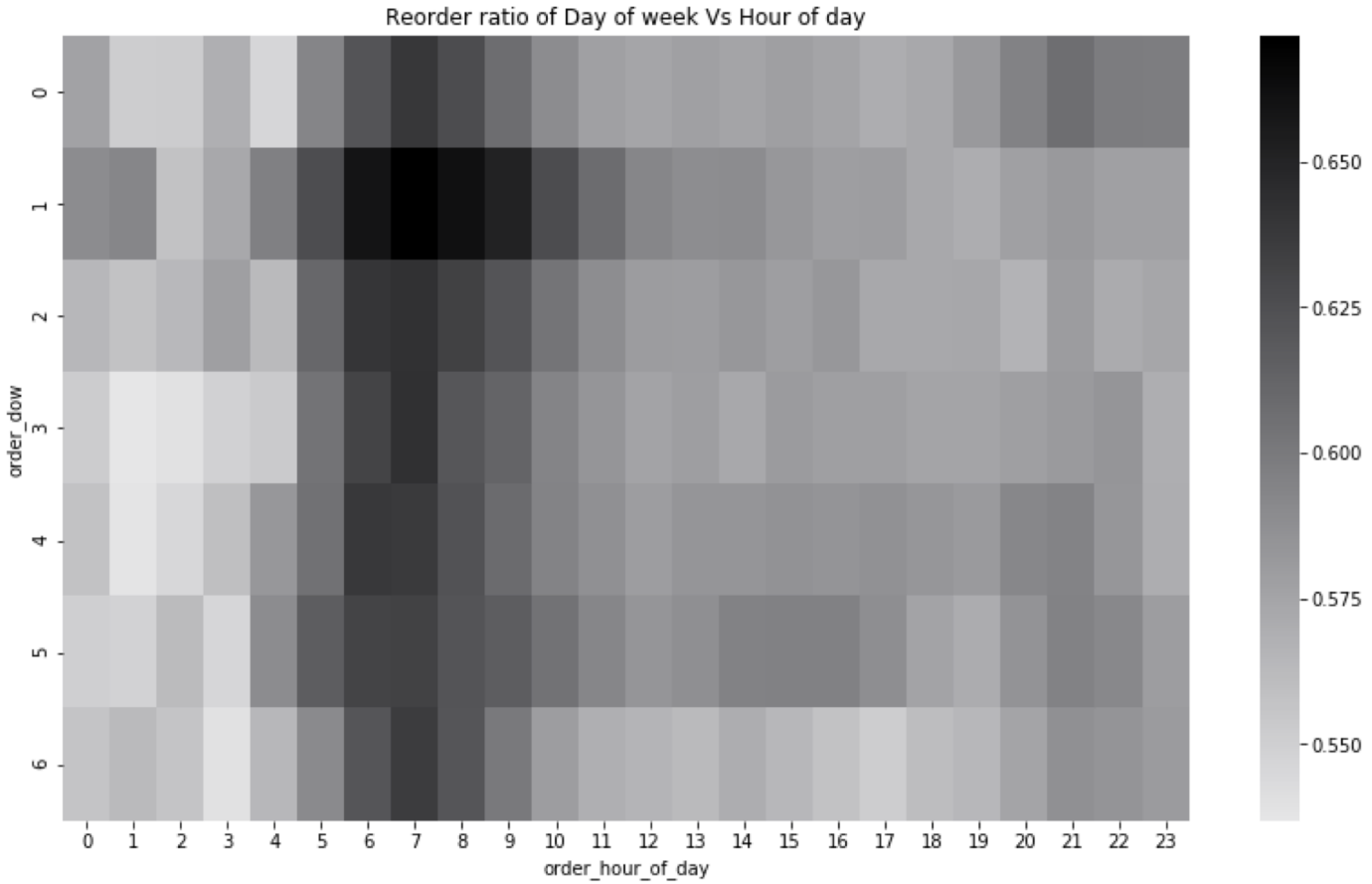


In terms of products in the basket, 1-10 products are the most common basket size

EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis



Heat map showing the time of the day and day of the week where most of the reordering occurs

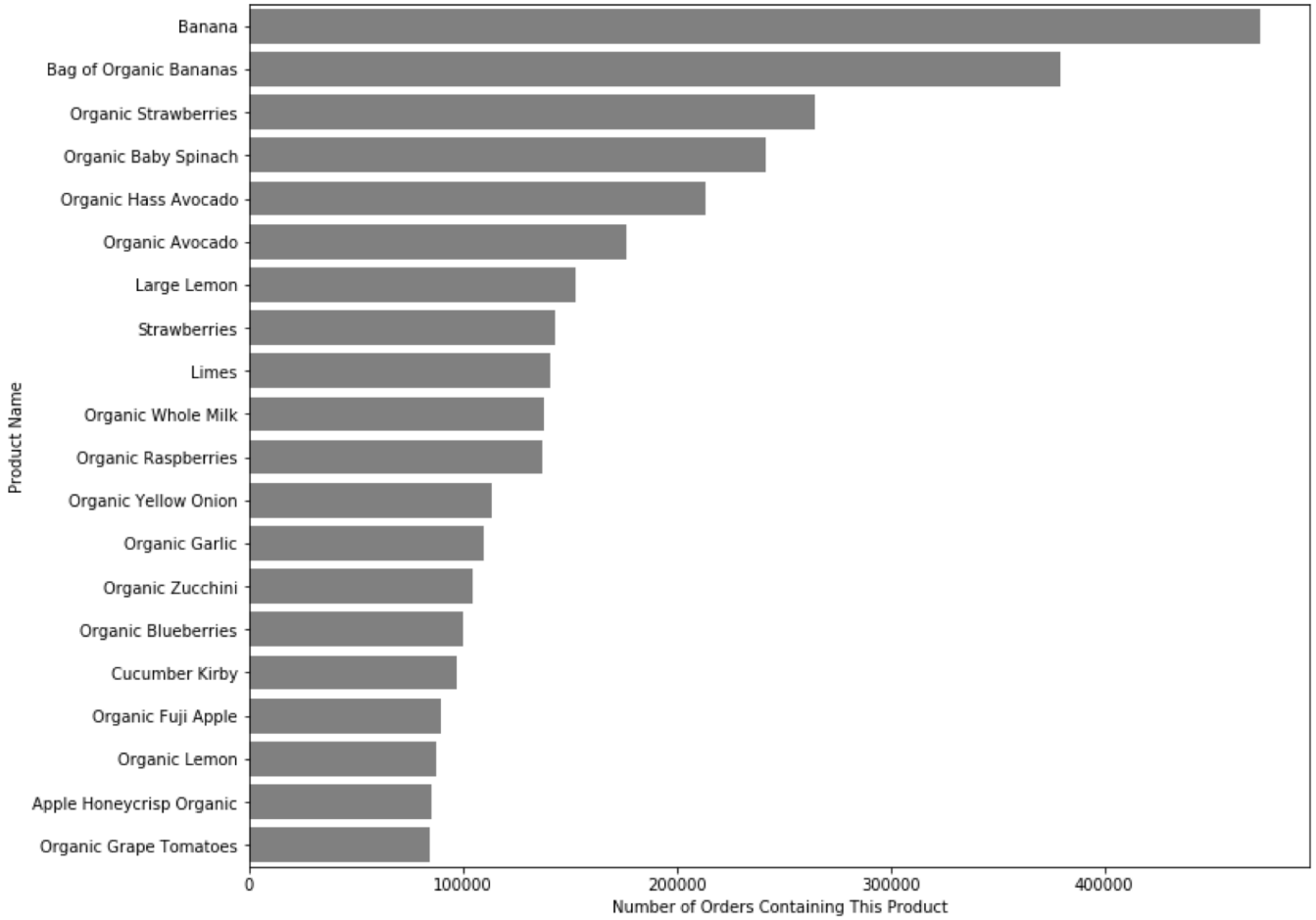
The darker the map, the strongest the ratio thus the likelihood of reordering happening

In general, morning are the moments of the day where the reordering is stronger.

EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis



Bananas are the most popular product across the dataset

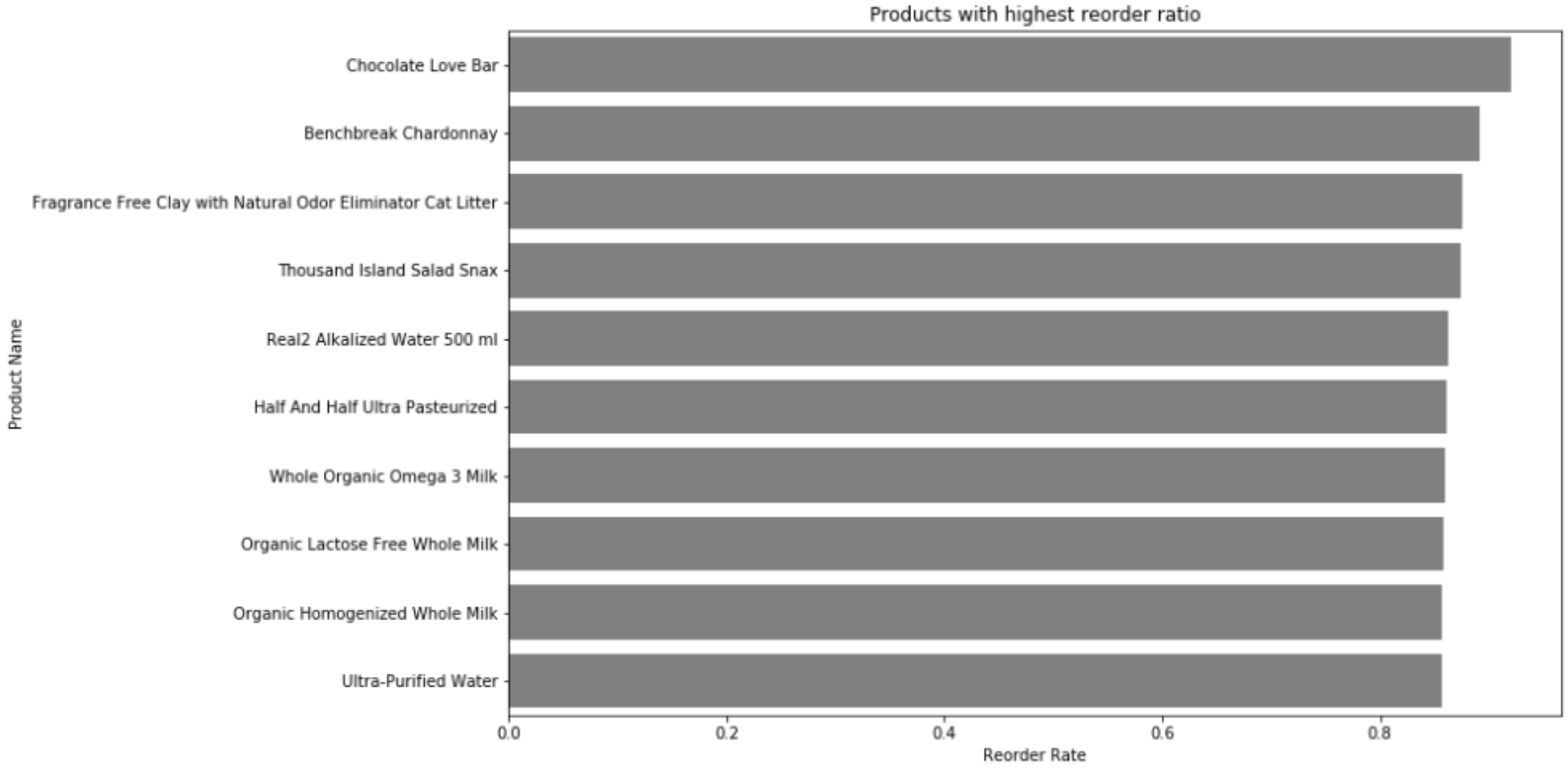
Followed by strawberries, baby spinach and avocado

In general, the most popular ordered products are fresh produce

EPFL EXTENSION SCHOOL

instacart

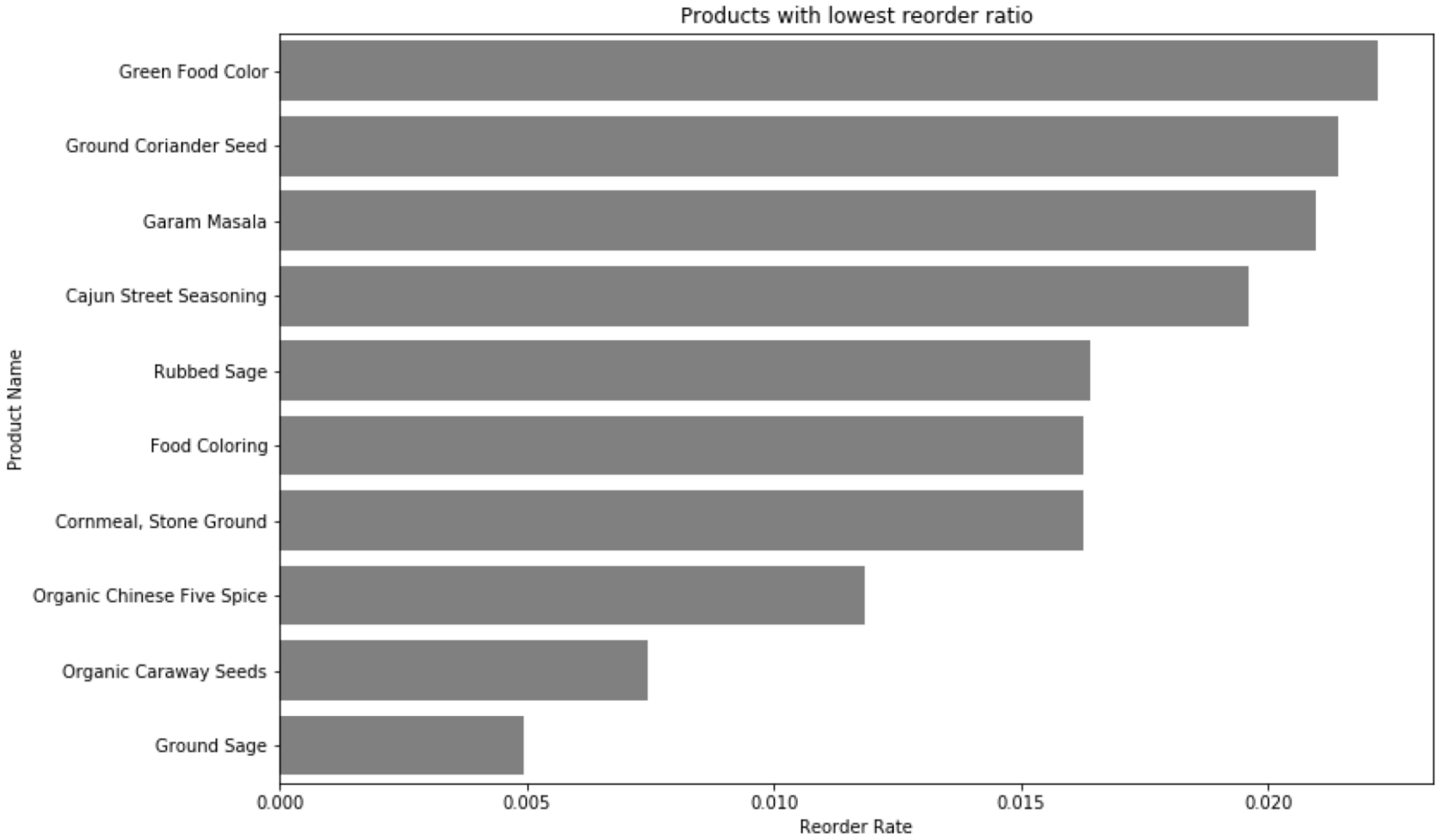
Exploratory and Descriptive Analysis



EPFL EXTENSION SCHOOL

instacart

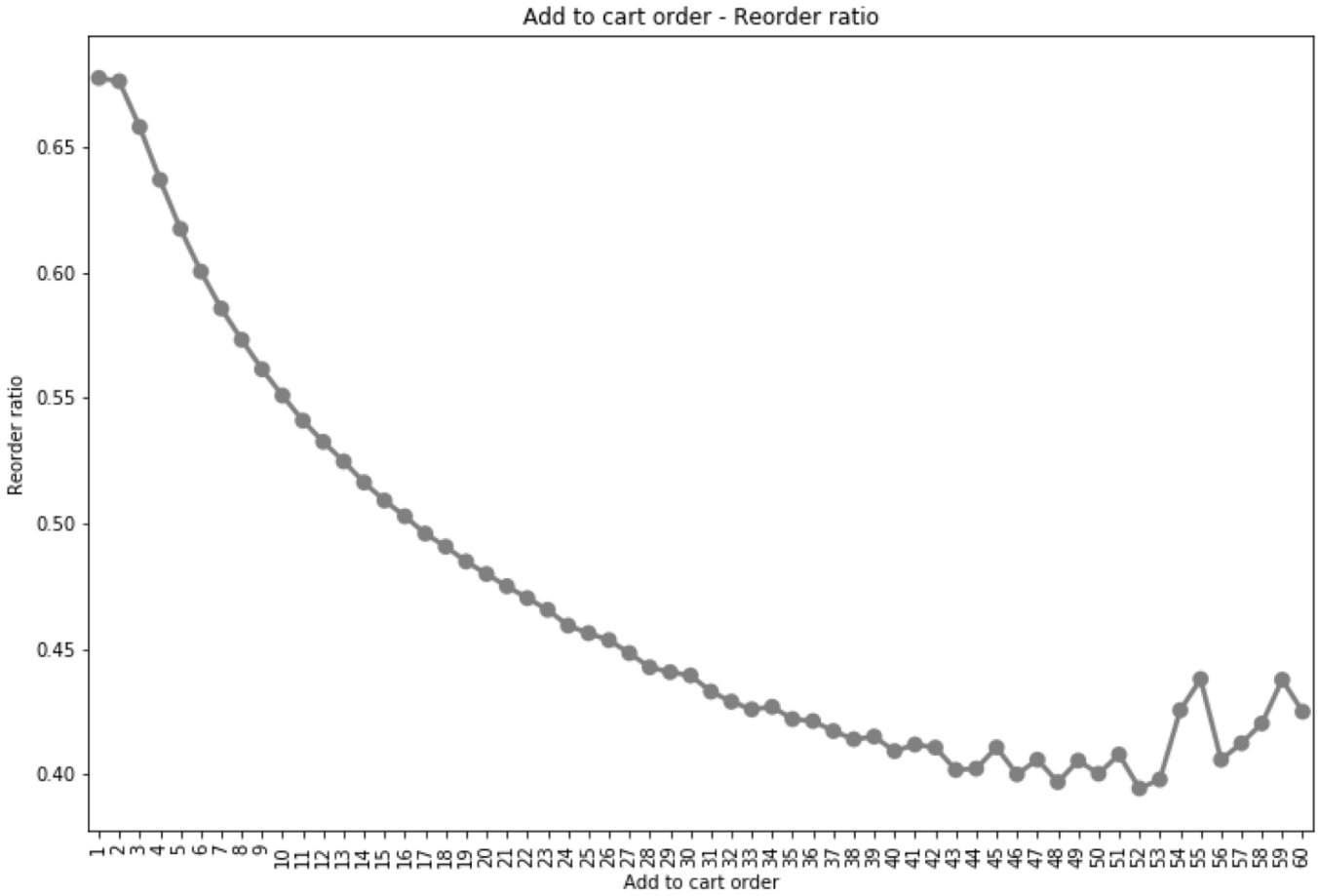
Exploratory and Descriptive Analysis



EPFL EXTENSION SCHOOL

instacart

Exploratory and Descriptive Analysis



The sooner a user includes a certain product in the cart, more likely it is to order it again on the next time

EPFL EXTENSION SCHOOL

instacart

Machine Learning, a first approach

- Sampling data due to computing restraints (1000, 10’000, 100’000)
- Models used:
  - Baseline
  - Stochastic Gradient Classifier
  - K-Nearest Neighbours
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Neural Networks with Tensor Flow
- Used Pipeline with Standard Scaler combined with Grid Search Cross Validation

EPFL EXTENSION SCHOOL

instacart

Baseline (n = 1000)

- It acts as a reference point and is the simplest possible prediction
- Does not depend on the patterns that appear in the data
- Estimator that classifies all the time as 1, in our case, user reorders Product

Baseline

```
from sklearn.base import BaseEstimator
from sklearn.model_selection import cross_val_score

class reorder(BaseEstimator):
    def fit(self, X, y=None):
        pass
    def predict(self, X):
        return np.ones((len(X),1), dtype=bool)

start_time1 = datetime.now()
reorder_clf = reorder()
end_time1 = datetime.now()
print ('Accuracy: {:.3f}'.format(cross_val_score(reorder_clf, X_train, y_train, cv=3, scoring = 'accuracy').sum()/3))
print('\nTime to run in seconds {}\n'.format((end_time1-start_time1).total_seconds()))
```

```
Accuracy: 0.585
```

EPFL EXTENSION SCHOOL

instacart

Stochastic Gradient Descent Classifier (SGD) (n = 1000)

- Advantage of being capable of handling very large datasets efficiently
- SGD deals with training instances independently, one at a time

Stochastic Gradient Classifier

```
from sklearn.linear_model import SGDClassifier

start_time1 = datetime.now()

# Create the estimator
sgd_clf = SGDClassifier(random_state=42)

# Fit the model to train data
sgd_clf.fit(X_train, y_train)

# Accuracy on test set
accuracy_sgd = sgd_clf.score(X_test, y_test)

end_time1 = datetime.now()
print('Accuracy: {:.3f}'.format(accuracy_sgd))
print('\nTime to run in seconds {}\n'.format((end_time1-start_time1).total_seconds()))
```

```
Accuracy: 0.630
```

EPFL EXTENSION SCHOOL

instacart

K Nearest Neighbours (n = 1000)

- *k*-NN is an intuitive algorithm that works for both regression and classification tasks
- It uses a distance metric and a decision function that can be adapted to the problem
- Accuracy of 0.62 using the standard algorithm
- Increased accuracy to 0.69, tuning k and distance metric using Grid Search

EPFL EXTENSION SCHOOL

instacart

Logistic Regression (n = 1000)

- Minimizes the cross-entropy (CE) loss function
- Accuracy of 0.63 using the standard algorithm from Scikit learn
- Using Grid Search combined with pipeline and standardization:
  - C – controls the weight of the loss function

$$C * \text{loss} + \text{penalization}$$

  - Using ”saga” as the solver, a variant of gradient descent
  - Improved accuracy to 0.67

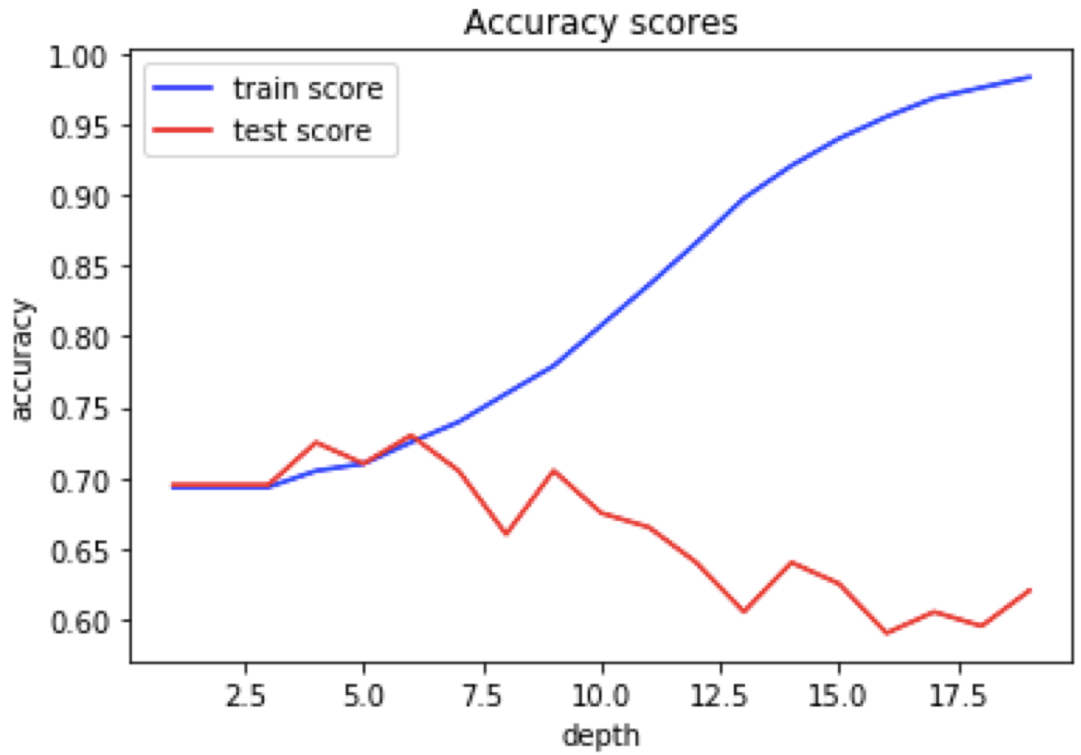
```
# Create pipeline
pipe = Pipeline([
    ('scaler', None),
    ('logreg',LogisticRegression())
])

# Create cross-validation object
grid_cv=GridSearchCV(pipe,[{
    'scaler': [StandardScaler()],
    'logreg__C':[0.0001,0.1,1000],
    'logreg__solver':['saga']
}],cv=3)

# Fit estimator
grid_cv.fit(X_train,y_train)
```

Decision Tree (n = 1000)

- The idea behind decision trees is to learn a set of if-then-else rules that lead to a final decision. In classification tasks, this decision is a label
- Optimal depth at 6
- Accuracy of 0.73



EPFL EXTENSION SCHOOL

instacart

Random Forest (n = 1000)

- Combination of decision trees into a random forest classifier
- Best accuracy at 0.75
- Using 250 ensemble of trees

	n estimators	test accuracy
0	50	0.690
1	100	0.715
2	150	0.720
3	200	0.715
4	250	0.750
5	300	0.750
6	350	0.720
7	400	0.725

EPFL EXTENSION SCHOOL

instacart

Machine Learning, a first approach

- Sampling of 1000, 10’000 and 100’000 because using the whole dataset was impossible to compute
- Accuracy was calculated for each combination of:
  - Sample and computing time
- Decision Trees and Random Forest seem to perform better
- Computing time increases significantly with n = 100’000

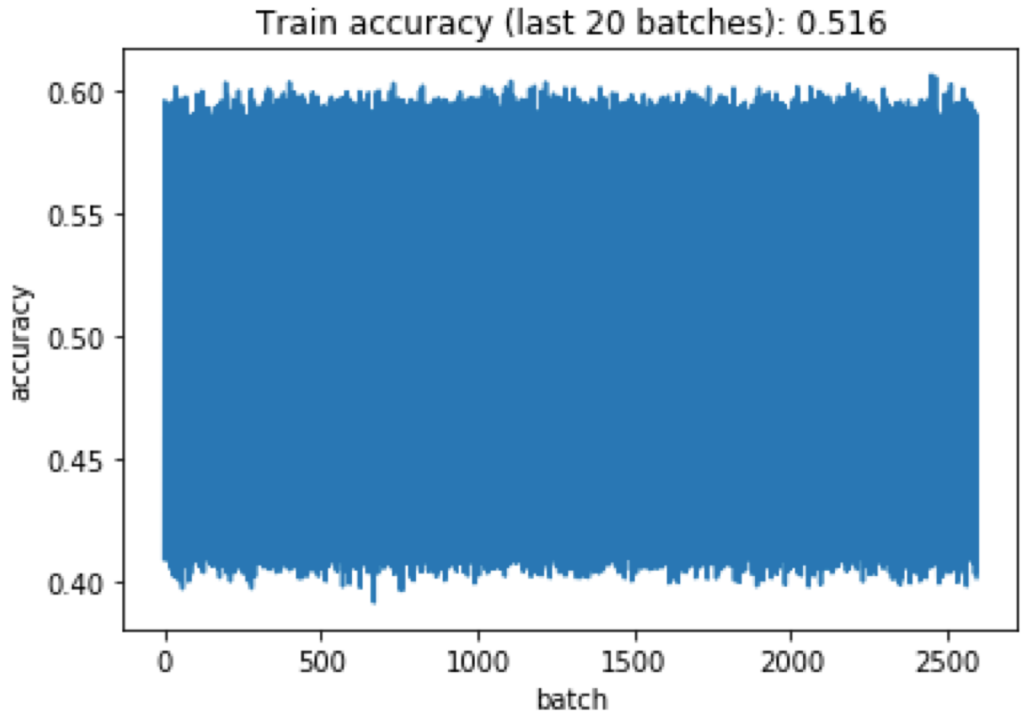
	_model	accuracy	sample	time to run
0	Baseline	0.614997	1000	00:00:00.000140
1	SGD	0.200000	1000	00:00:00.010032
2	KNN	0.750000	1000	00:00:00.006132
3	Logistic Regression	0.750000	1000	00:00:00.108472
4	Decision Tree	0.850000	1000	00:00:00.107971
5	Random Forest	0.750000	1000	00:00:04.798598
6	Baseline	0.584874	10000	00:00:00.000052
7	SGD	0.630000	10000	00:00:00.007604
8	KNN	0.690000	10000	00:00:00.057443
9	Logistic Regression	0.690000	10000	00:00:00.392436
10	Decision Tree	0.730000	10000	00:00:00.863500
11	Random Forest	0.750000	10000	00:00:37.065603
12	Baseline	0.591400	100000	00:00:00.000072
13	SGD	0.419500	100000	00:00:00.107132
14	KNN	0.704500	100000	00:00:07.664451
15	Logistic Regression	0.698000	100000	00:00:04.432800
16	Decision Tree	0.726500	100000	00:00:10.094575
17	Random Forest	0.722000	100000	00:06:59.754093

EPFL EXTENSION SCHOOL

instacart

Neural Network with Tensor Flow

- Solution allows to use the data in batches of 10’000
- Limited performance (accuracy around 0.60)



EPFL EXTENSION SCHOOL

instacart

Machine Learning, second approach

- Features Engineering
- Models used:
  - Baseline
  - Stochastic Gradient Classifier
  - K-Nearest Neighbours
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - Neural Networks with Tensor Flow
- Process

EPFL EXTENSION SCHOOL

instacart

Features Engineering

- Engineer a new data table from the existing tables
- The index of the new table:
  - Created a list of unique pairs of users and products from the prior set of orders. Gave this list the label "up_pair" for "user-product pair". This list was used as a unique index of our new data table
- The target variable:
  - Created a binary target variable for wether or not a user reordered a product
  - If a user brought a product in the prior set of orders and reordered the product in the train set of orders, then the target variable is assigned the value 1
  - If a user bought a product in the prior set, however they did not reorder the product in the train set, then the target variable is assigned the value 0

EPFL EXTENSION SCHOOL

instacart

Features Engineering

- Four explanatory variables. These features are as follows:
- Given the user-product pair of (User A, Product X):
  1. Total_buy_n5: the total number of times User A bought Product X out of the most recent orders
  2. Order_ratio_by_chance_n5: the proportion of User A's most recent orders in which User A had the chance to buy X, and indeed did so. Here, a "chance" refers to the number of opportunities the user had for buying the item after first encountering (viz., buying) it.
  3. Useritem_order_days_max_n5: the longest number of days User A has recently gone without buying Product X. We are only considering the 5 most recent orders
  4. Useritem_order_days_min_n5: the shortest number of days that User A has recently gone without buying Product X. Again, only considering the 5 most recent orders

EPFL EXTENSION SCHOOL

instacart

Machine Learning, second approach

- SGD classifier – accuracy of 0.93
- Logistic Regression – accuracy of 0.94
- Decision Tree – accuracy of 0.93
- Random Forest – accuracy of 0.94
- Neural Networks with Tensor Flow – 0.92

EPFL EXTENSION SCHOOL

instacart

Conclusion

- Feature engineering is the most important art in machine learning which creates the huge difference between a good model and a bad model
- It is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data
- It requires domain knowledge along with a pinch of creativity to extract more information without adding more data.
- Although feature engineering takes a considerable amount of time of the overall ML pipeline, the benefits far outweigh the efforts and time.

EPFL EXTENSION SCHOOL

instacart

Observations

- During the project, several iterations have been made due to the size of the dataset after joining/concatenating data either with Pandas or with SQL
- A Google Cloud Platform was created in order to overcome this problem as this allowed to increase significantly the computing power
- Going forward, in my short term learning objectives Apache Spark might be interesting to tackle big data issues.