**Build Your Machine Learning Project**

In this capstone project, you will leverage what you’ve learned throughout the program to build a machine learning engineer project of your choosing. Your project deliverables are:

1. A Github repository of your work.
2. A blog post written for a technical audience, or a deployed web application powered by data.

In this capstone project, you will leverage what you’ve learned throughout the Nanodegree program to solve a problem of your choice.

1. You will first **define** the problem you want to solve and investigate potential solutions.
2. Next, you will **analyze** the problem through visualizations and data exploration to have a better understanding of what algorithms and features are appropriate for solving it.
3. You will then **implement** your algorithms and metrics of choice, documenting the preprocessing, refinement, and post-processing steps along the way.
4. Afterwards, you will collect **results** about your findings, visualize significant quantities, validate/justify your results, and make any concluding remarks about whether your implementation adequately solves the problem.
5. Finally, you will **construct** a blog post to document all of the steps from start to finish of your project, or deploy your results into a web application.

**Setting Yourself Apart**

An important part of landing a job or advancing your career as a machine learning engineer is setting yourself apart through impressive projects. By now, you've completed several guided projects, and now's your chance to show off your skills and creativity. You'll receive a review and feedback from a Udacity mentor, and they will focus on how your project demonstrates your skills.

This project is designed to prepare you for delivering a polished, end-to-end solution report of a real-world problem in a field of interest. When developing new technology, or deriving adaptations of previous technology, properly documenting your process is critical for both validating and replicating your results.

Things you will learn by completing this project:

* How to research and investigate a real-world problem of interest.
* How to accurately apply specific data science algorithms and techniques.
* How to properly analyze and visualize your data and results for validity.
* How to document and write a report of your work.

## Software Requirements

**Your project must be written in Python 3.x**. Given the free-form nature of the capstone, the software and libraries you will need to successfully complete your work will vary depending on the chosen application area and problem definition. Because of this, it is imperative that all necessary software and libraries used in your capstone project are accessible to the reviewer and clearly documented. Information regarding the software and libraries your project makes use of should be included in the README along with your submission. Please note that proprietary software, software that requires private licenses, or software behind a paywall or login account should be avoided.

## Data Requirements

Every capstone project will most certainly require some form of dataset or input data structure (input text files, images, etc.). Similar to the software requirements above, the data you use must either be publicly accessible or provided by you during the submission process, and private or proprietary data should not be used without expressed permission. Please take into consideration the file size of your data — while there is no strict upper limit, input files that are excessively large may require reviewers longer than an acceptable amount of time to acquire all of your project files and/or execute the provided development code. This can take away from the reviewer's time that could be put towards evaluating your submission. If the data you are working with fits the criteria of being too large, consider whether you can work with a subset of the data instead, or provide a representative sample of the data which the reviewer may use to verify the solution explored in the project.

## Ethics

Udacity's A/B Testing course has a segment that discusses [the sensitivity of data](https://classroom.udacity.com/courses/ud257/lessons/3998098714/concepts/39997087540923) (free course link) and the expectation of privacy from those whose information has been collected. While most data you find available to the public will not have any ethical complications, it is extremely important that you are considering where the data you are using came from, and whether that data contains any sensitive information. For example, if you worked for a bank and wanted to use customers' bank statements as part of your project, this would most likely be an unethical choice of data and should be avoided.

NEXT

### Selecting a Project

Think about a technical field or domain that you are passionate about, such as robotics, virtual reality, finance, natural language processing, or even artificial intelligence (the possibilities are endless!). Then, choose an existing problem within that domain that you are interested in which you could solve by applying machine learning techniques. Be sure that you have collected all of the resources needed (such as data sets) to complete this project, and make the appropriate citations wherever necessary in Github (and your blog if that is the path you decide to pursue). Below are a few suggested problem areas you could explore if you are unsure what your passion is:

* [Robot Motion Planning](https://docs.google.com/document/d/1ZFCH6jS3A5At7_v5IUM5OpAXJYiutFuSIjTzV_E-vdE/pub)
* [Healthcare](https://docs.google.com/document/d/e/2PACX-1vT_MjTWMs1-668wgKALIK61AUHbqbKhp4UuUvYK99u0vc_R0K68Cqq5Br0Nf0aoGmg8XVAyfjd1MT8s/pub)
* [Computer Vision](https://docs.google.com/document/d/1y-XfjkPFgUQxFIQ9bBncUSjs4HOf5E-45FrLYNBsZb4/pub)
* [Education](https://docs.google.com/document/d/1vjerjRQnWs1kLbZagDYT6rNqiwAG23Yj45oUY88IAxI/pub)
* [Investment and Trading](https://docs.google.com/document/d/1ycGeb1QYKATG6jvz74SAMqxrlek9Ed4RYrzWNhWS-0Q/pub)

In addition, you may find a technical domain (along with the problem and dataset) as competitions on platforms such as [Kaggle](http://kaggle.com/), or [Devpost](http://devpost.com/). This can be helpful for discovering a particular problem you may be interested in solving as an alternative to the suggested problem areas above. In many cases, some of the requirements for the capstone project are already defined for you when choosing from these platforms.

#### Udacity Specific Projects

* **Customer Segmentation Report for Arvato Financial Services**
* **Optimizing App Offers With Starbucks**
* **Use Convolutional Neural Networks to Identify Dog Breeds**  
  (If you decide to do complete this project, you can find additional content to assist in the extra-curricular portion of this program).

#### Check out Sample Projects

* Here are two projects that can give you an idea of what a final blog post might look like. Each of these meets the requirements for the capstone project: [project 1](https://github.com/udacity/machine-learning/blob/master/projects/capstone/report-example-1.pdf) and [project 2](https://github.com/udacity/machine-learning/blob/master/projects/capstone/report-example-3.pdf).

No matter what project you decide to complete, you will want to make sure to check the [project rubric here](https://review.udacity.com/#!/rubrics/2345/view).

NEXT

# Create a Customer Segmentation Report for Arvato Financial Solutions

To introduce yourself to the scenario you'll be investigating in this capstone project option, take a look at the following video with Timo Reis from Arvato Financial Solutions.

## Steps to Complete This Project

The project has three major steps: the customer segmentation report, the supervised learning model, and the Kaggle Competition.

#### 1. Customer Segmentation Report

This section will be similar to the corresponding project in Term 1 of the program, but the datasets now include more features that you can potentially use. You'll begin the project by using unsupervised learning methods to analyze attributes of established customers and the general population in order to create customer segments.

#### 2. Supervised Learning Model

You'll have access to a third dataset with attributes from targets of a mail order campaign. You'll use the previous analysis to build a machine learning model that predicts whether or not each individual will respond to the campaign.

#### 3. Kaggle Competition

Once you've chosen a model, you'll use it to make predictions on the campaign data as part of a Kaggle Competition. You'll rank the individuals by how likely they are to convert to being a customer, and see how your modeling skills measure up against your fellow students.

NEXT

**Terms & Conditions**

In addition to Udacity's Terms of Use and other policies, your downloading and use of the **AZ Direct GmbH** data solely for use in the **Unsupervised Learning** and **Bertelsmann Capstone** projects are governed by the following additional terms and conditions. The big takeaways:

1. You agree to **AZ Direct GmbH's** General Terms provided below and that you only have the right to download and use the **AZ Direct GmbH** data solely to complete the data mining task which is part of the **Unsupervised Learning** and **Bertelsmann Capstone** projects for the Udacity Data Science Nanodegree program.
2. You are prohibited from using the **AZ Direct GmbH** data in any other context.
3. You are also required and hereby represent and warrant that you will delete any and all data you downloaded within 2 weeks after your completion of the **Unsupervised Learning** and **Bertelsmann Capstone** projects and the program.
4. If you do not agree to these additional terms, you will not be allowed to access the data for this project.

The full terms are provided in the workspace below. You will then be asked in the next workspace to agree to these terms before gaining access to the project, which you may also choose to download if you would like to read in full the terms.

These same exact terms are provided in the next workspace, where you will be asked to accept the terms prior to gaining access to the data.

# Starbuck's Capstone Challenge

Instructions for the project can be found in the Starbucks Project Workspace.

## Dataset overview

* The program used to create the data simulates how people make purchasing decisions and how those decisions are influenced by promotional offers.
* Each person in the simulation has some hidden traits that influence their purchasing patterns and are associated with their observable traits. People produce various events, including receiving offers, opening offers, and making purchases.
* As a simplification, there are no explicit products to track. Only the amounts of each transaction or offer are recorded.
* There are three types of offers that can be sent: buy-one-get-one (BOGO), discount, and informational. In a BOGO offer, a user needs to spend a certain amount to get a reward equal to that threshold amount. In a discount, a user gains a reward equal to a fraction of the amount spent. In an informational offer, there is no reward, but neither is there a requisite amount that the user is expected to spend. Offers can be delivered via multiple channels.
* The basic task is to use the data to identify which groups of people are most responsive to each type of offer, and how best to present each type of offer.

## Data Dictionary

### profile.json

Rewards program users (17000 users x 5 fields)

* gender: (categorical) M, F, O, or null
* age: (numeric) missing value encoded as 118
* id: (string/hash)
* became\_member\_on: (date) format YYYYMMDD
* income: (numeric)

### portfolio.json

Offers sent during 30-day test period (10 offers x 6 fields)

* reward: (numeric) money awarded for the amount spent
* channels: (list) web, email, mobile, social
* difficulty: (numeric) money required to be spent to receive reward
* duration: (numeric) time for offer to be open, in days
* offer\_type: (string) bogo, discount, informational
* id: (string/hash)

### transcript.json

Event log (306648 events x 4 fields)

* person: (string/hash)
* event: (string) offer received, offer viewed, transaction, offer completed
* value: (dictionary) different values depending on event type
  + offer id: (string/hash) not associated with any "transaction"
  + amount: (numeric) money spent in "transaction"
  + reward: (numeric) money gained from "offer completed"
* time: (numeric) hours after start of test

# Starbucks Capstone Challenge

### Introduction

This data set contains simulated data that mimics customer behavior on the Starbucks rewards mobile app. Once every few days, Starbucks sends out an offer to users of the mobile app. An offer can be merely an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). Some users might not receive any offer during certain weeks.

Not all users receive the same offer, and that is the challenge to solve with this data set.

Your task is to combine transaction, demographic and offer data to determine which demographic groups respond best to which offer type. This data set is a simplified version of the real Starbucks app because the underlying simulator only has one product whereas Starbucks actually sells dozens of products.

Every offer has a validity period before the offer expires. As an example, a BOGO offer might be valid for only 5 days. You'll see in the data set that informational offers have a validity period even though these ads are merely providing information about a product; for example, if an informational offer has 7 days of validity, you can assume the customer is feeling the influence of the offer for 7 days after receiving the advertisement.

You'll be given transactional data showing user purchases made on the app including the timestamp of purchase and the amount of money spent on a purchase. This transactional data also has a record for each offer that a user receives as well as a record for when a user actually views the offer. There are also records for when a user completes an offer.

Keep in mind as well that someone using the app might make a purchase through the app without having received an offer or seen an offer.

### Example

To give an example, a user could receive a discount offer buy 10 dollars get 2 off on Monday. The offer is valid for 10 days from receipt. If the customer accumulates at least 10 dollars in purchases during the validity period, the customer completes the offer.

However, there are a few things to watch out for in this data set. Customers do not opt into the offers that they receive; in other words, a user can receive an offer, never actually view the offer, and still complete the offer. For example, a user might receive the "buy 10 dollars get 2 dollars off offer", but the user never opens the offer during the 10 day validity period. The customer spends 15 dollars during those ten days. There will be an offer completion record in the data set; however, the customer was not influenced by the offer because the customer never viewed the offer.

### Cleaning

This makes data cleaning especially important and tricky.

You'll also want to take into account that some demographic groups will make purchases even if they don't receive an offer. From a business perspective, if a customer is going to make a 10 dollar purchase without an offer anyway, you wouldn't want to send a buy 10 dollars get 2 dollars off offer. You'll want to try to assess what a certain demographic group will buy when not receiving any offers.

### Final Advice

Because this is a capstone project, you are free to analyze the data any way you see fit. For example, you could build a machine learning model that predicts how much someone will spend based on demographics and offer type. Or you could build a model that predicts whether or not someone will respond to an offer. Or, you don't need to build a machine learning model at all. You could develop a set of heuristics that determine what offer you should send to each customer (i.e., 75 percent of women customers who were 35 years old responded to offer A vs 40 percent from the same demographic to offer B, so send offer A).

# Data Sets

The data is contained in three files:

* portfolio.json - containing offer ids and meta data about each offer (duration, type, etc.)
* profile.json - demographic data for each customer
* transcript.json - records for transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

**portfolio.json**

* id (string) - offer id
* offer\_type (string) - type of offer ie BOGO, discount, informational
* difficulty (int) - minimum required spend to complete an offer
* reward (int) - reward given for completing an offer
* duration (int) - time for offer to be open, in days
* channels (list of strings)

**profile.json**

* age (int) - age of the customer
* became\_member\_on (int) - date when customer created an app account
* gender (str) - gender of the customer (note some entries contain 'O' for other rather than M or F)
* id (str) - customer id
* income (float) - customer's income

**transcript.json**

* event (str) - record description (ie transaction, offer received, offer viewed, etc.)
* person (str) - customer id
* time (int) - time in hours since start of test. The data begins at time t=0
* value - (dict of strings) - either an offer id or transaction amount depending on the record

**Note:** If you are using the workspace, you will need to go to the terminal and run the command conda update pandas before reading in the files. This is because the version of pandas in the workspace cannot read in the transcript.json file correctly, but the newest version of pandas can. You can access the termnal from the orange icon in the top left of this notebook.

You can see how to access the terminal and how the install works using the two images below. First you need to access the terminal:

Then you will want to run the above command:

Finally, when you enter back into the notebook (use the jupyter icon again), you should be able to run the below cell without any errors.

In [1]:



**import** pandas **as** pd

**import** numpy **as** np

**import** math

**import** json

**%** matplotlib inline

​

*# read in the json files*

portfolio **=** pd.read\_json('data/portfolio.json', orient**=**'records', lines**=True**)

profile **=** pd.read\_json('data/profile.json', orient**=**'records', lines**=True**)

transcript **=** pd.read\_json('data/transcript.json', orient**=**'records', lines**=True**)

In [ ]:



​

## **Project Submission**

**Have project questions?** Ask a technical mentor or search for existing answers!

ASK A MENTOR

###### DUE DATE

**Jan 28**

###### STATUS

**Not submitted**

Due at: Thu, Jan 28 9:17 am

In this capstone project proposal, prior to completing the following **Capstone Project**, you will leverage what you’ve learned throughout the Nanodegree program to author a proposal for solving a problem of your choice by applying machine learning algorithms and techniques. A project proposal encompasses seven key points:

* The project's **domain background** — the field of research where the project is derived;
* A **problem statement** — a problem being investigated for which a solution will be defined;
* The **datasets and inputs** — data or inputs being used for the problem;
* A **solution statement** — the solution proposed for the problem given;
* A **benchmark model** — some simple or historical model or result to compare the defined solution to;
* A set of **evaluation metrics** — functional representations for how the solution can be measured;
* An outline of the **project design** — how the solution will be developed and results obtained.

Think about a technical field or domain that you are passionate about, such as robotics, virtual reality, finance, natural language processing, or even artificial intelligence (the possibilities are endless!). Then, choose an existing problem within that domain that you are interested in which you could solve by applying machine learning algorithms and techniques. Be sure that you have collected all of the resources needed (such as datasets, inputs, and research) to complete this project, and make the appropriate citations wherever necessary in your proposal.

In addition, you may find a technical domain (along with the problem and dataset) as competitions on platforms such as [**Kaggle**](http://kaggle.com/), or [**Devpost**](http://devpost.com/). This can be helpful for discovering a particular problem you may be interested in solving as an alternative to the suggested problem areas above. In many cases, some of the requirements for the capstone proposal are already defined for you when choosing from these platforms.

### Evaluation

Your project will be reviewed by a Udacity reviewer against the [**Capstone Project Proposal rubric**](https://review.udacity.com/#!/rubrics/410/view). Be sure to review this rubric thoroughly and self-evaluate your project before submission. All criteria found in the rubric must be meeting specifications for you to pass.

### Submission Files

At a minimum, your submission will be required to have the following files listed below. If your submission method of choice is uploading an archive (\*.zip), please take into consideration the total file size. You will need to include

* A project proposal, in PDF format only, with the name **proposal.pdf**, addressing each of the seven key points of a proposal. The recommended page length for a proposal is approximately two to three pages.
* Any additional supporting material such as datasets, images, or input files that are necessary for your project and proposal. If these files are too large and you are uploading your submission, instead provide appropriate means of acquiring the necessary files in an included README.md file.

Once you have collected these files and reviewed the project rubric, proceed to the project submission page.

### I'm Ready!

When you're ready to submit your project, click on the **Submit Project** button at the bottom of the page.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at [**\*\*machine-support@udacity.com**](mailto:**machine-support@udacity.com)\*\* or visit us in the [**discussion forums**](http://discussions.udacity.com/).

### What's Next?

You will get an email as soon as your reviewer has feedback for you. In the meantime, review your next project and feel free to get started on it or the courses supporting it!