UNIVERSITY OF SAN FRANCISCO | College of Arts and Sciences

MS in Data Science

ASSESSMENT REPORT
ACADEMIC YEAR 2017 – 2018

Academic Effectiveness Annual Assessment Resource Page:

https://myusf.usfca.edu/arts-sciences/faculty-resources/academic-effectiveness/assessment

Email to submit the report: assessment_cas@usfca.edu

Important: Please write the name of your program or department in the subject line.

For example: FineArts_Major (if you decide to submit a separate report for major and minor); FineArts_Aggregate (when submitting an aggregate report)

I. LOGISTICS & PROGRAM LEARNING OUTCOMES

1. **Please indicate the name and email of the program contact person to whom feedback should be sent (usually Chair, Program Director, or Faculty Assessment Coordinator).**

Terence Parr, parrt@cs.usfca.edu

2. **Were any changes made to the program mission statement since the last assessment cycle in October 2017? Kindly state "Yes" or "No." Please provide the current mission statement below. If you are submitting an aggregate report, please provide the current mission statements of both the major and the minor program.**

No

3. **Were any changes made to the program learning outcomes (PLOs) since the last assessment cycle in October 2017? Kindly state "Yes" or "No." Please provide the current PLOs below. If you are submitting an aggregate report, please provide the current PLOs for both the major and the minor programs.**

Note: Major revisions in the program learning outcomes need to go through the College Curriculum Committee (contact: Professor Joshua Gamson, gamson@usfca.edu). Minor editorial changes are not required to go through the College Curriculum Committee.

Yes, here are the new PLO ratified by a vote of faculty members over email on March 23, 2018:

(1) Program Learning Outcome 1. Possess a theoretical understanding of classical statistical models (e.g., generalized linear models, linear time series models, etc.), as well as the ability to apply those models effectively.

(2) Program Learning Outcome 2. Possess a theoretical understanding of machine learning techniques (e.g., random forests, neutral networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively of data.

(3) Program Learning Outcome 3. Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data.

(4) Program Learning Outcome 4. Be prepared for careers as data scientists by solving real-world data-driven business problems with other data scientists.

(5) Program Learning Outcome 5. Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.

The map of PLO to courses appears below.

4. Which particular Program Learning Outcome(s) did you assess for the academic year 2017-2018?

We assessed 2 through 4 (All PLOs provided at the end of this document):

(2) Possess a theoretical understanding of machine learning techniques (e.g., random forests, neutral networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively;
(3) Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data;
(4) Be prepared for careers as data scientists by solving real-world data-driven business problems with other data scientists

II. METHODOLOGY

5. Describe the methodology that you used to assess the PLO(s).

For example, "the department used questions that were inputted in the final examination pertaining directly to the <said PLO>. An independent group of faculty (not teaching the course) then evaluated the responses to the questions and gave the students a grade for responses to those questions."

Five data science faculty contributed representative questions from one or more written exams taken from six data science courses related to the three learning outcomes mentioned above. These learning outcomes are loosely described as computing–related. We decided on a scoring mechanism that could be used across all courses and all questions on those exams used in this report:

3. Student has mastered material necessary to answer a specific question
2. Student did not give a perfect answer to the question but has a solid grasp on the concepts
1. Student did not achieve a minimum level of competency for a specific question
0. Student did not answer

For the purposes of this assessment report, faculty returned to their exams, sometimes months after initially grading the students, to select and score assessment questions as 3, 2, or 1. All scoring data is attached, as are the exam questions themselves.

Here are the number of questions broken down by program learning objectives and course:

PLO	Questions in category
PLO2	10
PLO3	27
PLO4	9

Course	PLO2	PLO3	PLO4
MSAN621	5	0	0
MSAN630	5	0	1
MSAN691	0	3	3
MSAN692	0	5	3
MSAN694	0	9	1
MSAN697	0	10	1

All exam questions used in the creation of this assessment report are provided at the end of this document.

III. RESULTS & MAJOR FINDINGS

1. **What are the major takeaways from your assessment exercise?**

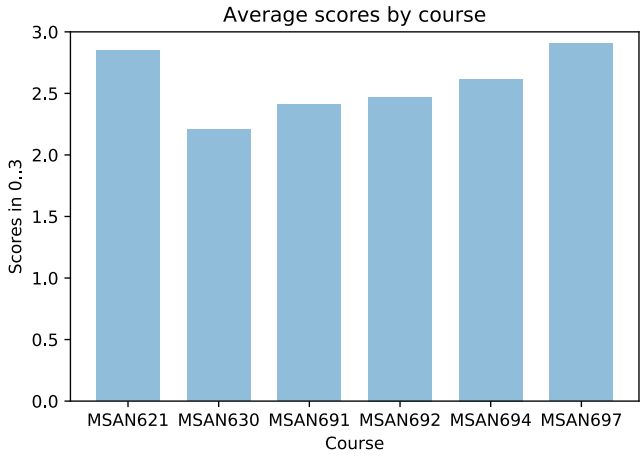
This section is for you to highlight the results of the exercise. Pertinent information here would include:

a. how well students mastered the outcome at the level they were intended to,

b. any trends noticed over the past few assessment cycles, and

c. the levels at which students mastered the outcome based on the rubric used.

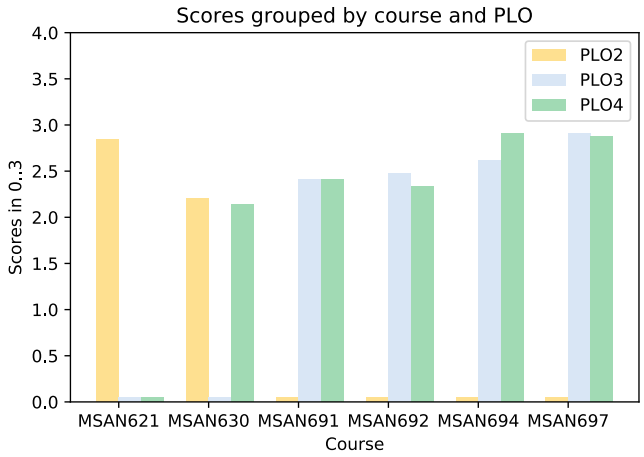
After compiling all of the data from six courses, we found some variation in score levels across courses but the overall average score for a single question was 2.58 out of 3. According to our score metric listed above, therefore, students achieve above "solid grasp on the concepts," but below complete mastery when we aggregate all questions from all courses. Here's a graph that shows the average score per question broken down by course:



(Please note that the number of questions per course varied significantly and so the average scores must be weighted appropriately to reach the correct 2.58 average).

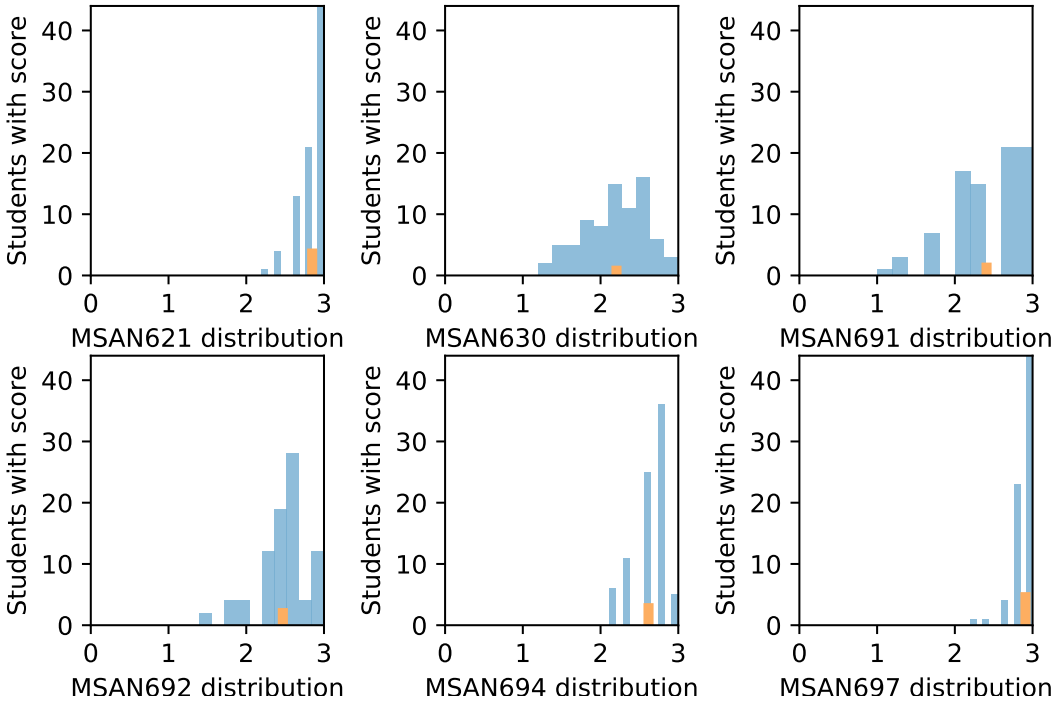
Courses MSAN694 and MSAN697 show improvement over the earlier courses MSAN691 and MSAN692. MSAN621 was taught by an adjunct and graded cooperatively between the instructor and two full-time faculty and so it is harder to directly compare those scores with the others, as they were taught and evaluated by the same full-time faculty member.

Here is a graph that breaks down the average scores by program learning objectives:



(As described above, not all learning objectives are covered in each class and so some of the bars are 0.)

The scores across learning objectives are very consistent within the same course, which indicates that the faculty member explicitly or implicitly instructed the students consistently across objectives. Breaking down the scores even further, here's a distribution (histogram) showing how many students got what average score per class (class average shown as orange rectangle):



At the opposite end of the spectrum, we see that in MSAN621 students master the course material for the most part. The courses are somewhere between “solid grasp” and “mastery”. Here are average scores for each learning objective, computed across all courses and all students:

PLO	Average score per question across courses
PLO2	2.53
PLO3	2.67
PLO4	2.46

Given the difficulty of our graduate program, we consider 2.53 to be an excellent outcome for PLO2 (theoretical understanding of machine learning techniques). PLO3’s 2.67 average score is even higher and shows a solid grasp of the material, heading towards mastery. This PLO deals with distributed computing, which is typically very far outside of the students’ prior experience (and the score is also reduced because of the blanks). We might have a bit more work to do for PLO4 (prepared for careers as data scientists by solving real-world data-driven business problems) as it is slightly lower at 2.46, but still shows solid grasp of the material.

IV. CLOSING THE LOOP

2. **Based on your results, what changes/modifications are you planning in order to achieve the desired level of mastery in the assessed learning outcome? This section could also address more long-term planning that your department/program is considering and does not require that any changes need to be implemented in the next academic year itself.**

Based upon these results, the data science faculty will discuss at the next faculty meeting and make recommendations to faculty about areas and topics to focus on. We might also recommend changing our scoring mechanism from 1..3 to 1..4 to get finer granularity on our metrics. We are discussing new and innovative forms of direct assessment on exams going forward. For example, we'd love to have students write small programs and have them automatically executed and tested as part of an online examination.

3. **What were the most important suggestions/feedback from the FDCD on your last assessment report (for academic year 2016-2017, submitted in October 2017)? How did you incorporate or address the suggestion(s) in this report?**

I believe that, due to the lateness of our previous report, we did not receive feedback.

ADDITIONAL MATERIALS

(Any rubrics used for assessment, relevant tables, charts and figures should be included here)

Program Learning outcomes (PLOs)

Upon successfully completing the Master of Science in Data Science (MSDS) program, our graduates will:

(1) Possess a theoretical understanding of classical statistical models (e.g., generalized linear models, linear time series models, etc.), as well as the ability to apply those models effectively;
(2) Possess a theoretical understanding of machine learning techniques (e.g., random forests, neutral networks, naive Bayes, k-means, etc.), as well as the ability to apply those techniques effectively;
(3) Effectively use modern programming languages (e.g., R, Python, SQL, etc.) and technologies (AWS, Hive, Spark, Hadoop, etc.) to scrape, clean, organize, query, summarize, visualize, and model large volumes and varieties of data;
(4) Be prepared for careers as data scientists by solving real-world data-driven business problems with other data scientists; and
(5) Develop professional communication skills (e.g., presentations, interviews, email etiquette, etc.), and begin integrating with the Bay Area data science community.

The latest version of the MSDS program learning outcomes was ratified by a vote of faculty members over email on March 23, 2018.

PLO course mapping

Program Phase	Bootcamp				Fall Module One					Fall Module Two					Intersession	Spring Module One					Spring Module Two					Summer 2			
	MSAN 501: Computation for Analytics	**MSAN 502:** Review of Linear Algebra	**MSAN 504:** Review of Probability and Statistics	**MSAN 593:** Exploratory Data Analysis	**MSAN 601:** Linear Regression Analysis	**MSAN 692:** Data Acquisition	**MSAN 610:** Communications for Analytics	**MSAN 691:** Relational Databases	**MSAN 640:** Seminar Series I	**MSAN 621:** Introduction to Machine Learning	**MSAN 604:** Time Series Analysis	**MSAN 694:** Distributed Computing	**MSAN 605:** Practicum I	**MSAN 641:** Seminar Series II	**MSAN 697:** Distributed Data Systems	**MSAN 628:** Computational Statistics	**MSAN 630:** Advanced Machine Learning	**MSAN 689:** Problem Solving with Python	**MSAN 625:** Practicum II	**MSAN 642:** Seminar Series III	**MSAN 622:** Data Visualization	**MSAN 629:** Experiments in Data Science	**MSAN 603:** Product Analytics	**MSAN 627:** Practicum III	**MSAN 643:** Seminar Series IV	**MSAN 631:** Special Topics in Analytics	**MSAN 627:** Practicum IV	**MSAN 644:** Seminar Series V	**10 required hours of Interview training**
Program Phase	**Bootcamp**				**Fall Module One**					**Fall Module Two**					**Intersession**	**Spring Module One**					**Spring Module Two**					**Summer 2**			
Units	1	1	1	1	2	2	1	1	0	2	2	1	1	0	2	2	2	1	2	0	2	2	2	2	0	2	1	0	0
PLO1			I	I	I						D		I			D			D			M		M		M	M		M
PLO2				I	I					D	I		I		D	M	M	D	D				D	D		M	M		
PLO3	I	I		I		D		D		D	I	M	I		D	D	M	D	D		D		M	D			M		
PLO4				I		I	I	I					D					D	D			D	D	M			M		M
PLO5							I		I				I	D					D	M			D	M	M		M	M	D
# PLOs supported by each course	1	1	1	4	2	2	2	2	1	2	3	1	5	1	2	3	2	3	5	1	1	2	4	5	1	2	5	1	3

3. (1 points) What part of an HTTP web page fetch request sends URL parameters.

5. (1 points) Give a single terminal command that fetches http://foo.com/bar and also prints out the HTTP protocol "conversation" with the web server program at foo.com.

10. (1 points) Imagine you have launched an instance at Amazon Web services and started a server program listening at port 8080. The server program responds using curl from the command line on the server but curl does not work from your laptop using the public IP address. What is the most likely problem?

15. (2 points) What two primary advantages does Selenium have over requests/BeautifulSoup?

18. (3 points) You are given a nonempty list of URLs, urls, and an accompanying list of page titles, titles. Write Python code to create a list of dict objects, one per URL/title pair. Call the list data. Each dict must have two keys, url and title, with the appropriate values. (Note: I'm not asking for a single dict mapping each url to its title or vice versa.)

MSAN 697 Distributed Data Systems exams

Quiz 1

Question 3 1 pts

Question 5 1 pts

```
file_name = "../README.md"
lines = sc.textFile(file_name)
word = lines.flatMap(lambda line : line.split())
word_map = word.map(lambda word : (word,1))
word_map_reduce = word_map.reduceByKey(lambda a,b : a+b)
finalRDD = word_map_reduce.collect()
```

Considering RDD lineage, how many stages does this code have?

Quiz 2

Question 2 — 1 pts

createDataFrame( ) allows to specify schema.

- ○ True
- ○ False

Question 4 — 1 pts

When **df_1, df_2, and output** are given as below, what is the query that generates the output?

```
df_1
+---+---+
| id|age|
+---+---+
|  a|  2|
|  b|  2|
|  c|  3|
+---+---+


df_2
+---+----+
| id|name|
+---+----+
|  a|Mary|
|  b|Mike|
|  c| Kim|
|  d|Abby|
+---+----+
```

```
ouput
+---+----+----+
| id| age|name|
+---+----+----+
|  d|null|Abby|
|  c|   3| Kim|
|  b|   2|Mike|
|  a|   2|Mary|
+---+----+----+
```


- ○ df_1.join(df_2, 'id', 'rightOuter')
- ○ df_1.join(df_2, 'name', 'rightOuter')
- ○ df_1.join(df_2, 'id', 'leftOuter').show()
- ○ df_1.join(df_2, 'id').show()

Quiz 3

Question 1 1 pts

Which function would you use to randomly split a data set into a training and test data set to build a classification model?

- randomSplit()
- evaluate()
- transform()
- fit()

Question 3 1 pts

For training a logistic regression model with a data frame called adulttrain, which method should be called in the following code?

ca.instructure.com/courses/1574686/quizzes/2320685/take?preview=1

Quiz: Quiz 3 - Spark ML (Closed-book)

```
#Train the model.
from pyspark.ml.classification import LogisticRegression
lr = LogisticRegression(regParam=0.01, maxIter=1000, fitIntercept=Tr
lrmodel = lr.[ ](adulttrain)
```

- evaluator()
- transform()
- fit()
- ParamGridBuilder()

Question 4 1 pts

For Spark ML, some popular machine learning algorithms are not included, because the algorithms are not designed for parallel platforms.

- True
- False

Question 5 1 pts

Which method should be called in the following code?

```
from pyspark.ml.feature import StringIndexer
si = StringIndexer(inputCol=c, outputCol=c+"-num")
sm = si.fit(newdf)
newdf = sm.___(newdf).drop(c)
```

- evaluator()
- transform()
- fit()

- pipeline()

Quiz 4

Question 1 1 pts

Postgres does not have native mechanisms to partition the database across a cluster of nodes.

- True
- False

Question 2 1 pts

For NoSQL, which of the following is false?

- Mostly developed in the 21st century.
- Mostly open-source databases.
- Mostly not using SQL as a query language.
- Strictly satisfying concurrency control (ACID).

Question 1 0.5 pts

For RDDs, which of the following are false?

- It's an acronym for "Resilient Distributed Dataset".
- If a node fails, it recovers data from other replica nodes by default.
- Even if data is distributed, RDDs abstract away their distributed nature.
- Once created, RDDs are read-only.

Question 3 0.5 pts

```
words = sc.parallelize(["MSAN694", "MSAN694 Distributed Computing"])
```

```
words.map(lambda x : x.split()).count()
```

What is the output of the last line, ***words.map(lambda x : x.split()).count()***?

- 2
- 3
- 4
- None of the above

Question 6 0.5 pts

For Apache Spark, which of the following are false?

- Apache Spark supports streaming functionalities.
- Apache Spark is generally faster than Hadoop MapReduce.
- Apache Spark is not suitable for iterative algorithms.
- Using Apache Spark, you can write distributed programs.

Question 7 — 0.5 pts

sfca.instructure.com/courses/1574500/quizzes/2319456/take?preview=1

Quiz: Final (Closed-book)

```
def sum_1(x,y):
    return x+y
```

```
def sum_2(x):
    return sum(x)
```

```
def sum_3(x,y):
    z = x+y
```

```
nums = sc.parallelize([1,2,3,4])
```

Which of the following commands return the sum of ***nums***?

- [ ] nums.reduce(sum_1)
- [ ] nums.reduce(sum_2)
- [ ] nums.reduce(sum_3)
- [ ] None of the above

Question 8 — 0.5 pts

Which of the following operations triggers a computation to return the result, assuming a proper function is given as a parameter?

- ( ) map()
- ( ) filter()
- ( ) take()
- ( ) flatMap()

Question 15 1 pts

```
pair_rdd.collect()
```

```
[(1, 2), (1, 3), (2, 3), (2, 4)]
```

```
pair_rdd.combineByKey(0, lambda x,y : x+y, lambda x,y : x+y).collect()
```

The last line, ***pair_rdd.combineByKey(0, lambda x,y : x+y, lambda x,y : x+y).collect()***, returns ***[(1, 5), (2, 7)]***.

- True
- False

Question 19 1 pts

```
from pyspark import StorageLevel
```

```
pair_rdd.cache()
```

```
pair_rdd.persist(StorageLevel.MEMORY_AND_DISK)
```

The last line, ***pair_rdd.persist(StorageLevel.MEMORY_AND_DISK)*** throws an error.

- True
- False

Question 23 0.5 pts

The application above is a UI of YARN resource manager.

- True
- False

Question 25 0.5 pts

```
words = sc.parallelize(["Spark", "Spark is fun", "Spark is fun"], 6)
```

```
words.saveAsTextFile("File")
```

What is the output of the last line, ***words.saveAsTextFile("File")***?

- A folder called "File" which contains 3 files.
- A folder called "File" which contains 6 files.
- A file called "File".
- 6 files called "File_1", "File_2", ... , "File_6".

MSAN 691 Relational Databases exam

1. What are the top five salesperson (SalesID) in terms of *number* of items sold?

```
select SalesID
from sales
group by 1
order by sum(1) desc
limit 5;
```

11. Of all the salespeople from California ("CA") or Pennsylvania ("PA"), return the one (Name and SalesID) with the highest amount of revenue.

```
select
        SP.SalesID, Name
from
        SP
left join
        Sales
using( SalesID)
where SP.state in ('CA', 'PA')
group by 1,2
order by sum( Rev) desc
limit 1;
```

17. An item is called "complex" if the ratio of unique parts (PIDs) to total parts is more than .90% and is called "simple" if the ratio is below .25%. Write a query which returns one row with three columns: the first column should be the total revenue generated by complex items, the second should be the total revenue generated by simple items and the third should be the revenue generated by items which are neither simple nor complex.

```
select
        sum( case when rat > .9 then amt else 0 end ) as TR_complext
        , sum( case when rat < .25 then amt else 0 end) as TR_Simple
        , sum( case when rat <=.9 and rat >= .25 then amt else 0 end) as TR_neither
from
        (select IID, count(1)::float / sum( NoPart) as rat from Assembly
        group by 1) as lhs
left join
        Sales
using(IID)
```

1. (3 points) Given a forest of n tree estimators, $tree_i$, in a Random Forest (RF) regressor, give Python code or pseudo-code using elements like $tree_i$.predict(x) for single observation x to yield a RF scalar prediction. (Ignore that scikit predict() actually wants a matrix not vector.)

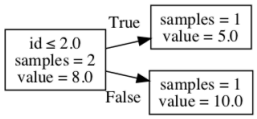
5. (1 points) With a RF regressor, the R^2 score is 0.83, the out-of-bag R^2 score is 0.80, and the validation R^2 score is 0.29. Circle which of the following most likely explains the situation.

1. The model is overfit
2. The model is underfit
3. The validation set is not like the training set

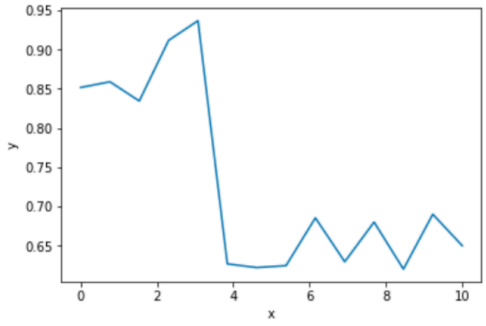
7. (**5 points**) Given the following data frame with predictor `id` and predicted (dependent) `price`, draw a single decision tree with a single split. Each leaf node must contain `samples=___` and `value=___`. Non-leaves must contain the split variable expression in addition such as `blort<=foo`. Edges must be labeled true or false.

id	price
1	5
3	10

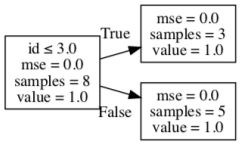
Solution:



9. (**2 points**) Given the following graph of independent variable x and dependent variable y, where is the best first split on x? Fill in the blank: `x<=`_____ 3.0



solution



10. (**5 points**) Consider computing the random forest importance value for a single variable held in vector `X`. The dependent, predicted variable is in vector `y`. Fill in this function with Python or pseudocode to compute the importance of `X` and return that importance. You can use functions like "shuffle(column name)" and "score(X,y)", for example. Shuffle shuffles all the values within a column of a data frame. Score takes a matrix/dataframe of observations and returns an R^2 score comparing results of prediction to truth y. You do not need to write a loop or list comprehension. *You may not retrain the random forest.*

```
def importanceA(rf, X, y):
    """
    For vector X with a single observation and predicted variable y, compute
    and return the important score for X in random forest model rf.
    """
```

1. Given training set $\{x_i, y_i\}_{i=1}^N$ and $x_i, y_i \in \mathbb{R}$. Compute the gradient descent equations for the following regression model. (Use the square error / loss).

$$y = a + b \cdot x + c \cdot x^2$$

2. Compute naive mean encoding of the feature "city".

city	y	city_mean_enc
San Francisco	1	
San Francisco	0	
San Francisco	1	
San Jose	1	
San Jose	1	
San Jose	0	
San Jose	0	

3. Given the following specification of a neural network. Compute prediction you will give to $x = (-1, -1)$. (Show me the computation.)

$b_1^{[1]}$	$w_{11}^{[1]}$	$w_{12}^{[1]}$	$b_2^{[1]}$	$w_{21}^{[1]}$	$w_{22}^{[1]}$	$b^{[2]}$	$w_1^{[2]}$	$w_2^{[2]}$
-30	20	20	10	-20	-20	-10	20	20

Assume that

$$\sigma(a) = h(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{otherwise.} \end{cases}$$

4. Consider the following dataset $(x^{(1)} = (0,0), y^{(1)} = 1), (x^{(2)} = (3,0), y^{(2)} = 1), (x^{(3)} = (1,1), y^{(3)} = -1)$ and $(x^{(4)} = (1,-1), y^{(4)} = 1)$ with corresponding weights $w_1 = \frac{3}{10}, w_2 = \frac{3}{10}, w_3 = \frac{3}{10}, w_4 = \frac{1}{10}$.

   (a) Compute the weighted misclassification rate of the following classifier. Which points are misclassified?

$$g(x) = \begin{cases} 1, & \text{if } x_1 > 0.5 \\ -1, & \text{otherwise} \end{cases}$$

   (b) Find a stump that minimizes the weighted misclassification rate. What is the best weighted misclassification rate?

5. You are tasked with solving an NLP classification problem. The input to your problem is a sentence. The output is 0 or 1. You have a training set of 2k sentences with labels. Your boss asked you to look into using pre-trained embeddings. Explain two strategies to transform your input sentences into training set with a **fixed** number of features.