Homework #1 **Due: turned in by Monday 01/27/2020 before class**

Carl Xi

(put your name above)

Total grade: _____ out of ___100___ points

There are 2 parts in this homework assignment with 17 numbered questions in total. Please answer them all and submit your assignment as a single PDF or Word file by uploading it to the HW1 drop-box on the course website.

Part I. Multiple Choice Questions (60 points; 4 points each)

1. **Which three are essential skills for a data scientist? (Choose three)**
   - **A. Software engineering**
   - B. System administration
   - **C. Statistics and mathematics**
   - **D. Domain experience**
   - E. Multidimensional data modeling
   - F. Database administration

2. **What are three common practices of data scientist working with big data? (Choose three)**
   - **A. Applying machine learning techniques**
   - **B. Working with large and disparate data**
   - C. Minimizing data redundancy through normalization
   - **D. Cleaning and transforming data**
   - E. Publishing findings in academic journals

3. **What is the first step in the lifecycle of a typical data science project?**
   - A. Identify the desired outcome
   - B. Capture the data
   - C. Determine which data is required
   - **D. Define the problem**

4. **Why should a data scientist iterate through the project lifecycle multiple times during one project? (Choose three)**
   - **A. To change techniques as the amount of data scales up**
   - **B. To benefit from lessons learned during the lifecycle**
   - C. To recover from a cluster node failure
   - **D. To incorporate new problems and new data discovered during the lifecycle**
   - E. Because the statistical significance of an inference did not reach the required threshold

5. **What are two reasons a data scientist conducts a preliminary analysis using a small amount of data? (Choose two)**
   - A. Because simple, smaller analyses are superior to complex, larger analyses
   - B. Because the results of large-scale analyses are difficult to communicate
   - **C. Because initial small-scale analyses can inform later large-scale analyses**
   - **D. Because a small-scale analysis can be used to validate an approach correct**

6. **When considering data quality and provenance, a data scientist should seek to acquire data that is: (Choose two)**
   - **A. Accurate**
   - B. Structured
   - C. Raw
   - **D. From a reputable source**

7. **Which are three valid reasons for anonymizing data? (Choose three)**
   A. Anonymization simplifies the data transformation process
   B. Laws, policies, and standards may require anonymization
   C. Anonymization mitigates the damage from intrusions
   D. Anonymization makes re-identification impossible
   E. Anonymization protects against legal liability in the event of an attack

8. **Which model typically yields the best predictions of future data?**
   A. A model that fits a general pattern observed in the data correct
   B. A model that exactly fits the observed data

9. **Which question can be answered using linear regression?**
   A. Based on its text content, what is the probability that a message is spam?
   B. What is the relationship between the monthly fees charged to a customer and the likelihood they cancel their subscription?
   C. What is the relationship between a customer's age and the number of minutes per day they spend using a service?
   D. What is the distribution of customer age?

10. **What is a supervised learning algorithm?**
   A. An algorithm that discovers structure in data where no formal structure exists
   B. An algorithm that must be monitored by the data scientist as it runs
   C. An algorithm that automatically determines its ideal behavior to maximize a measure of performance
   D. An algorithm that requires a training dataset with known labels correct

11. **Which should you consider when training supervised machine learning algorithms?**
   A. A binary classification algorithm must be trained on a dataset consisting exclusively of records with a positive value of the label
   B. An algorithm typically performs best when trained on a dataset consisting of only the most relevant attributes
   C. The accuracy of an algorithm typically degrades as the volume of training data increases
   D. A superior algorithm operating on a smaller training dataset is typically more accurate than a simpler algorithm operating on a larger training dataset

12. **Data suggests that the average listening session for Earcloud users is longer for women than for men. You are designing an experiment to try to confirm this.**
    Your experiment confirms with high confidence that the average Earcloud user session is longer for women than for men. In the terminology of hypothesis testing, what decision is made?
   A. Accept the null hypothesis
   B. Reject the null hypothesis
   C. Accept the alternative hypothesis
   D. Reject the alternative hypothesis

13. Before running an A/B test on your recommender system to confirm that the cancellation rate for women decreases when they are recommended longer playlists, you will first run an A/A test. What should the A/A test do?

A. Recommend longer playlists to both women and men
B. Recommend longer playlists to women and shorter playlists to men
C. Divide women into two groups and make no change to the recommended playlists in either group
D. Divide women into two groups and recommend longer playlists to one group

14. What is a best practice when deploying a change to a recommender system?

A. Deploy the change to all users in one step so results from multiple experimental groups are not intermingled in log files
B. Deploy the change to 50% of users so the experimental and control groups are of equal size
C. Deploy the change to a small proportion of users then ramp up in phases to limit possible negative impacts
D. Run the original and changed versions on alternating days to isolate negative impacts by day

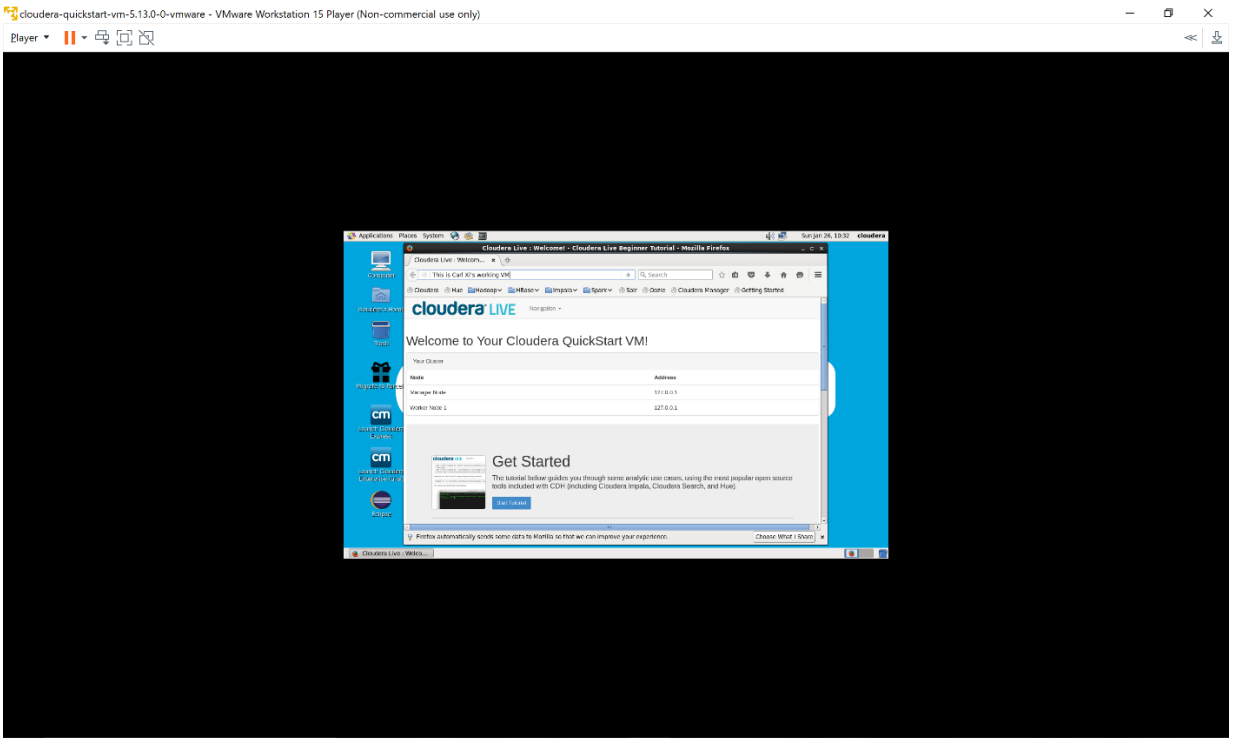
15. When is it most appropriate to use a lambda function in Python?

A. When the function will be used only once
B. When the function will be reused multiple times
C. When the function has no input parameters
D. When the function has no return value

Part II. Hands on (40 points)

1. Run the Cloudera QuickStart VM 5.13 (10 points)

For this question, you should download the Cloudera QuickStart VM (5.13 version) and run it on your local machine. The VM can be downloaded from here: https://www.cloudera.com/downloads/quickstart_vms/5-13.html Please provide a screenshot showing the VM running on your laptop.



2. Predictive Analytics Application (30 points)

Select a publicly available dataset (e.g., from the UC Irvine Machine Learning Repository, Kaggle, etc.) of your interest and then build and evaluate a predictive model of your choice using any tool \ programming language you prefer.

In general, use best practices when building and evaluating your models: optimize parameters on validation data, perform final evaluation on test data, etc.

As part of this question, you should present a brief overview of your predictive modeling process and discuss your results. Your overview should i) clearly describe the specific predictive task you selected and the class it belongs to (e.g., classification, clustering, etc.) as well as cover ii) why you selected the particular machine learning algorithm for the predictive task you selected. Make sure you also present iii) information about the model "goodness" (possible things to think about: confusion matrix, predictive accuracy, classification error, precision, recall, f-measure, ROC curves).

The presented overview should not exceed 2 pages.

Predictive Analytics Application

a) Problem Statement

This modeling assignment is inspired by two major problems with train ticket prices during the summer:

1. Ticket prices fluctuate unpredictably, people don't know what to expect for the route and times they want to travel.
2. Getting to the same destination might cost drastically less if rerouting is possible, but people often do not have knowledge of the prices of all possible routes.

b) Predictive Task

In one sentence: Can we predict the price of a train ticket during the summer months of April to October if we know when and where the passenger is traveling?

c) Data Description

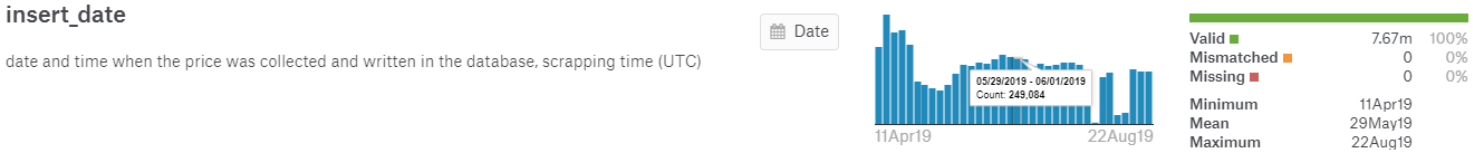
The dataset comes from Kaggle (Source: https://www.kaggle.com/thegurusteam/spanish-high-speed-rail-system-ticket-pricing) and is based on the Spanish High Speed Train Service (Renfe AVE) during the summer of 2019 (Apr. - Oct.). It contains 7.67 million rows and 9 columns).

d) Predictive Task Class

As the target variable is known and numeric (train ticket prices), this is clearly a supervised regression problem.

e) Data Cleaning

We created histograms for every variable and checked for any missing values



We first dropped all rows containing nulls reduced our dataset size from 7.67 million entries to 7.10 million entries. We then removed non-useful variables (like insert_date which only records when the entry is recorded into the system, not the timestamp of the actual ticket). We then dropped all duplicate entries and all non-AVE train entries in the dataset, which reduced our dataset to 89.515 entries and 70.,963 entries respectively.

f) Feature Engineering

We conducted feature engineering on numeric variables. We created year, month, day, hour, minute and day of the week categorical variables, an algorithm that checks whether departure/arrival date is a holiday and creates two binary variables. We then converted the origin and destination variables into "route" variable, which enables us to use fewer dummy variables while capturing the same information. We lastly created a travel_time variable representing the duration of the trip.

g) Data Transformation

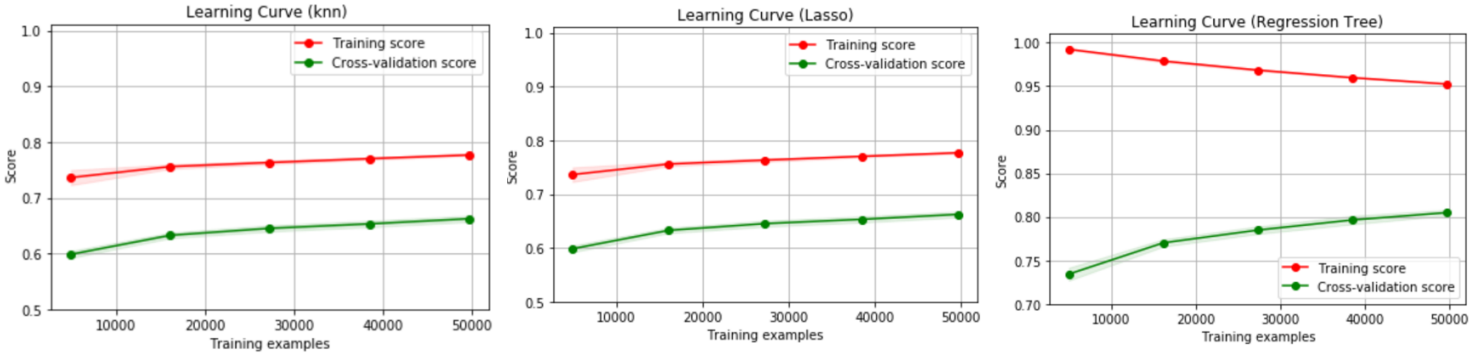
With all variables cleaned, and necessary variables created, we checked to see if any data transformation is needed. Luckily, most of our data is normally distributed with minimum skewness, so we did not have to conduct any transformation (e.g. log(variable), (variable)^2, etc.) Our final dataset contains 70,963 entries and 16 variables. As most of our variables are categorical, we ended up with 150 variables after we converted all categorical variables into dummy variables. Below is how our dataset looks like

```
result = pd.concat([AVE1, AVE3], axis=1)
result.head()
```

	train_type	price	train_class	fare	month	day	hour	minute	weekday	travel_time	month2	hour2	minute2	weekday2	route	Holiday	Holidayarr
0	AVE	85.1	Turista	Promo	4	12	5	50	Friday	185.0	4	8	55	Friday	1	0	0
1	AVE	73.1	Turista	Flexible	4	12	6	45	Friday	113.0	4	8	38	Friday	2	0	0
2	AVE	132.8	Preferente	Promo	4	12	7	0	Friday	150.0	4	9	30	Friday	1	0	0
3	AVE	76.3	Turista	Flexible	4	12	7	0	Friday	152.0	4	9	32	Friday	3	0	0
4	AVE	127.1	Turista	Flexible	4	12	7	0	Friday	150.0	4	9	30	Friday	1	0	0

h) Machine Learning Algorithm Selection

We used 3 modeling techniques (kNN, Regression Tree and Linear Regression) and compared their RMSE. We selected Lasso to represent linear regression as it outperformed ridge regression and automatically selects features, which helped us reduce the number of dimensions in our dataset (curse of dimensionally). Here are the resulting learning curves of the three techniques we chose.



i) Parameter Tuning/ Cross Validation

We conducted gridsearch, which tuned hyper parameters using Nested Cross Validation for the whole dataset. This helped us extract the best parameters based on finding the best estimator. We then conducted cross-validation to evaluate the generalization performance for the whole dataset

j) Performance Evaluation

We decided to use Regression Tree for our final model as it is robust to categorical variables (which our dataset has a lot of), captures nonlinear relationships better, performs better for big datasets (which ours is), and is easy to interpret. RMSE is chosen as the metric to evaluate our models because it gives us a interpretable value of the accuracy of our prediction, denoted in euros. In addition, since we did not transform our target variable, RMSE is comparable across all three models. Our results are as follows:

	MAE	MSE	RMSE
Lasso Regression	22.3	380.4	19.5€
Regression Tree	14.1	175.5	13.2€
k-Nearest Neighbors	18.6	280.1	16.7€
Naive	34.3	1040	32.2€

From the results, we can see that our model of choice (Regression Tree) performed the best, resulting in ticket price estimations within 13.2€ RMSE, especially considering that the average ticket is 74.63€ (calculated from our dataset). With 11.1 million AVE riders in 2019 alone and a 4.2% YoY growth rate, our model can help consumers save over 41.42 million EUR if they can use our model to find tickets that are just 5% cheaper than without our model.

Appendix:

Lasso Cross-Validation code in Jupyter Notebook:

```
cross_val_lasso = cross_val_score(Lasso(), X=X, y=y, cv=10 ,scoring = "neg_mean_squared_error")
print(" MSE: ", cross_val_lasso.mean(), " +/- ", cross_val_lasso.std())
```

```
 Performance:  -383.3880305419385  +/-  86.63509437159826
```

Post-Grid-Search Decision Tree Cross-Validation code in Jupyter Notebook:

```
GS_DT_CV = cross_val_score(GS_DT, X=X1, y=y1, cv=10)
print("Nested CV Performance: ",GS_DT_CV.mean(), " +/- ", GS_DT_CV.std())
```

```
Nested CV Performance:  -175.13154403181096  +/-  39.50586454444203
```

Grid-Search kNN Cross-Validation code in Jupyter Notebook:

```
GS_KNN = GridSearchCV(estimator=neighbors.KNeighborsRegressor(p=2,
                         metric='minkowski'),
                  param_grid=[{'n_neighbors': [1,3,5,7,9,11,13,15,17,19,21],
                               'weights':['uniform','distance']}],
                  scoring= "neg_mean_squared_error",
                  cv=5,
                  n_jobs=5)

GS_KNN.fit(X_std2,y2)

print("\n Parameter Tuning ")
print()
print("Non-nested CV f1: ", GS_KNN.best_score_)
print()
print("Optimal Parameter: ", GS_KNN.best_params_)
print()
print("Optimal Estimator: ", GS_KNN.best_estimator_) # Estimator that was chosen by the search, i.e. estimator which gave hig
print()
```

```
 Parameter Tuning

Non-nested CV f1:  -489.6189857815545

Optimal Parameter:  {'n_neighbors': 21, 'weights': 'distance'}

Optimal Estimator:  KNeighborsRegressor(algorithm='auto', leaf_size=30, metric='minkowski',
          metric_params=None, n_jobs=None, n_neighbors=21, p=2,
          weights='distance')
```