1 Python Basics

For help review Assignment 1 material.

1.1 File paths

What is the difference between a relative and absolute path? Give an example of each.

1.2 Data Structures

What type is returned with the following code?

$$df[[\text{"}age\text{"}]] \tag{1}$$

$$df[\text{"}age\text{"}] \tag{2}$$

1.3 Slicing

What is the difference between the following two lines of code?

$$df.iloc[0:1] \tag{3}$$

$$df.loc[0:1] \tag{4}$$

2 Machine Learning Fundamentals

For help review Assignment 2 material.

2.1 EDA

During exploratory data analysis, you create 2 histograms of a feature to show the distribution of the feature values in the training set, separated for positive (target=1) and negative (target=0) examples. You observe that the histograms for this particular feature look identical between the two target classes. Would the feature be useful for predicting the target class? Briefly justify your answer.

2.2 Text Processing

Why are free-text column features difficult to use in machine learning algorithms?

2.3 Cross-Validation

Below are the subscores from five folds of cross-validation for 2 machine learning models. Which model appears to be overfitting the most? Which model's average validation score would you trust the least? Which model would you choose to use and why?

Model A	test_score	train_score
0	0.68	0.99
1	0.63	0.99
2	0.70	0.99
3	0.72	0.99
4	0.60	0.99

Model B	test_score	train_score
0	0.76	0.91
1	0.82	0.90
2	0.85	0.90
3	0.79	0.90
4	0.80	0.91

3 Preprocessing

For help review Assignment 3 material.

3.1 Model Requirements

I just finished an initial exploratory data analysis of my dataset and split it into X_train, y_train, X_test, and y_test subsets. I tried to train sklearn's SVC model on X_train and y_train, but it didn't work. Give two reasons why it may have failed.

3.2 C Parameter

I'm tuning an RBF SVC model to pick the best hyperparameter value. I ran cross-validation on the SVC() model in sklearn with a variety of different C values to compare their mean test_score and train_scores. What should I consider when I choose the best C value? What effect does increasing the C value have on this model?

3.3 Transformations

You're tasked to train a machine learning model to predict whether a student in CPSC 330 will get an A+ on the final exam. You gather some preliminary information about the dataset.

```
df.head()
```

	enjoy	major	attnd	year	assign1	assign2	assign3	midterm	final
0	yes	CPSC	Excellent	3	92	93.0	84	92	A+
1	yes	MENG	Average	2	94	90.0	80	91	Not A+
2	yes	MATH	Poor	3	78	85.0	83	80	Not A+
3	no	MATH	Excellent	3	91	NaN	92	89	A+
4	yes	PSYC	Good	4	77	83.0	90	85	A+

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 370 entries, 0 to 369
Data Columns (total 10 columns):
```

```
train_df, test_df = train_test_split(df, test_size=0.2, random_state=42)
train_df['attnd'].unique()
```

#	Column	Non-Null Count	Dtype
0	enjoy	370 non-null	object
1	major	370 non-null	object
3	attnd	370 non-null	object
4	year	370 non-null	int64
5	assign1	370 non-null	int64
6	assign2	369 non-null	float64
7	assign3	370 non-null	int64
8	midterm	370 non-null	int64
9	final	370 non-null	object

```
array(['Excellent', 'Average', 'Poor', 'Good'],dtype=object)
```

Define a preprocessor ct which can be used to apply necessary transformations to the data. Your preprocessor will be used in cross-validation of an SVC() classifier model.

```
# Your code here
```

```
ct = # Your code here
pipe = make_pipeline(ct, SVC(random_state=42))
scores = cross_validate(pipe, X_train, y_train, return_train_score=True)
```

4 Hyperparameter Optimization

For help review Assignment 4 material.

4.1 GridSearchCV()

Two individuals are tuning a logistic regression model for text classification using count vectorizer. The goal is to find the optimal hyperparameter values for C and max_features.

Person A follows a manual approach, tuning one hyperparameter at a time. They first experiment with

different values of C while keeping max_features constant, and then vice versa. After a series of cross-validation calls, Person A selects the best hyperparameter values.

Person B, however, opts for a more automated approach and uses GridSearchCV to find the best hyperparameter values. Person B obtains different best values for C and max_features compared to Person A.

How is it possible that Person A and Person B ended up with different optimal hyperparameter values? Generally speaking, should their values agree? Why or why not?

4.2 Baselines

Describe a scenario where the absence of a baseline model might lead to misguided conclusions or suboptimal decisions in the model development process.

5 Performance & Interpretation

For help review Assignment 5 material.

5.1 F1 Score

Discuss a situation where a high F1 score might be more desirable than a high accuracy.

5.2 Interpreting Results

You're using the Ridge() model from sklearn to predict housing prices using the following features:

- Square Feet (eg. 1,800)
- Neighborhood Quality (categories = [poor, fair, good, excellent])
- House Style (categories = [ranch, colonial, split, cape]).

After training, your model learned the following coefficients:
ridge.intercept_ returns 16,000

sqft	250
neighborhood	25,000
style_ranch	5,000
style_colonial	25,000
style_split	15,000
style_cape	20,000

	sqft	neighborhood	style
A	1,200	poor	ranch

Part 1

Calculate the predicted price for House A given the following information:

Part 2

If House A were moved to an excellent neighborhood, how would that affect the predicted price?

Part 3

Say we had applied scaling to the numeric feature sqft before training our model. Would this affect our calculation of House A's predicted price from Part 1? Why or why not?

5.3 AUC

You created the following plot to visualize performance of three machine learning models (Model A, Model B, and Model C).

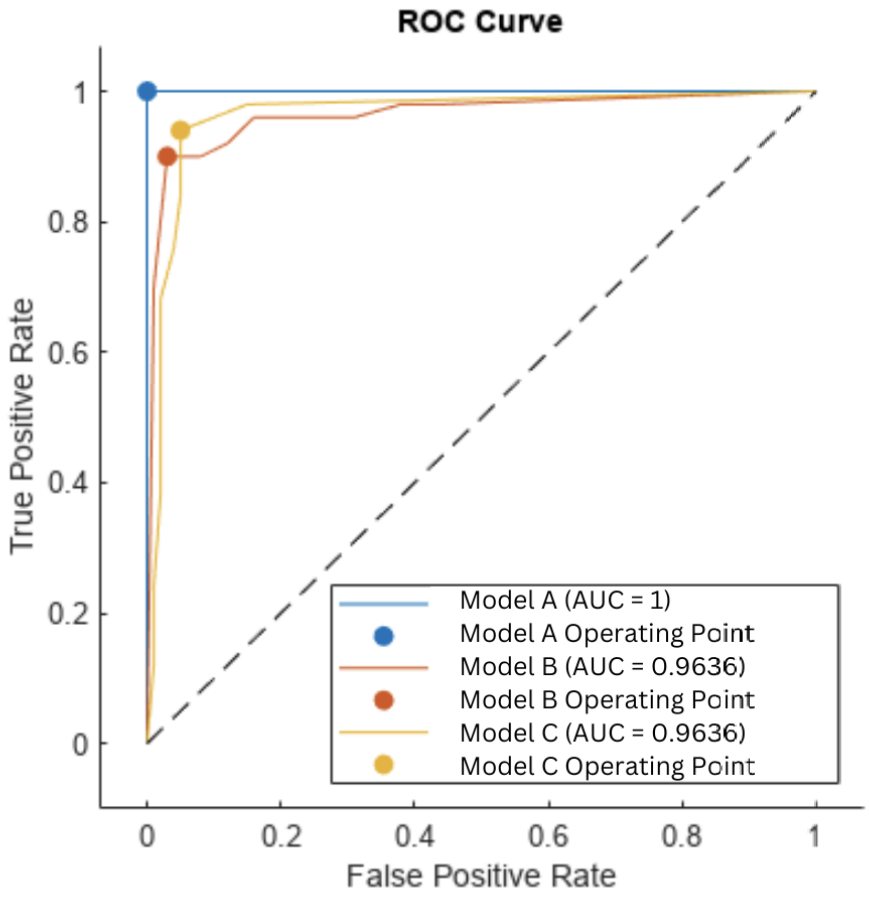
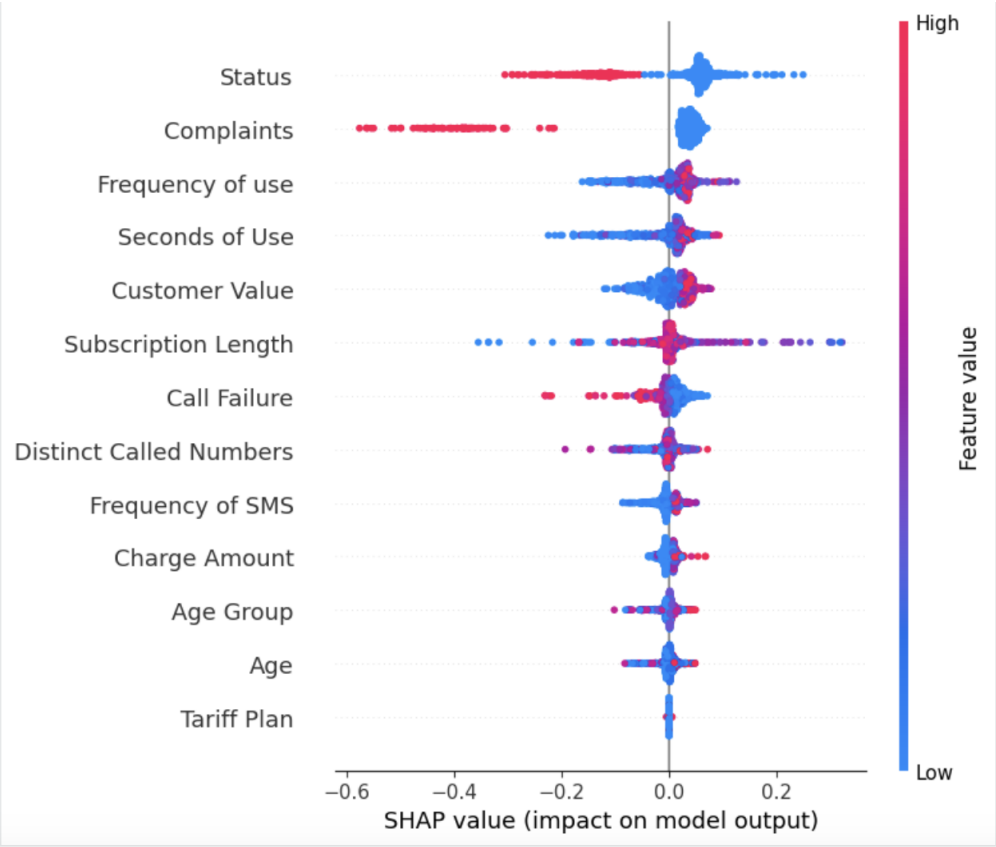


Figure 1: Note the operating point here is the deployed threshold.

Based on the information available in the plot, which model do you think is demonstrating the best overall performance? Does the Model B have a higher false positive rate than Model C? Briefly explain.

5.4 SHAP

Below is a SHAP summary plot for label 0 assigned by a RandomForestClassifier model trained on a telecommunication company's customer churn dataset:



Which feature is most important for the model output? Comment on the relationship of this feature and the target class 0.

6 Clustering

For help review Assignment 6 material.

6.1 DBSCAN

When might I want to use DBSCAN instead of K-Means for clustering?

6.2 eps & min_samples

Match the following combination of hyperparameter values and their impact on clustering.

a.	High eps, high min_samples	Results in larger, more loosely-defined clusters. Outliers more likely included.
b.	High eps, low min_samples	Leads to small, dense clusters. Algorithm tends to be less sensitive to noise
c.	Low eps, high min_samples	Tends to produce large, sparse clusters. Outliers are less likely to be included.
d.	Low eps, low min_samples	Tends to identify small, dense regions as clusters. Highly sensitive to noise.

7 Topic Modeling

For help review Assignment 7 material.

7.1 Text Pre-Processing

You're given a corpus of news articles and are tasked to identify high-level themes in the collection of documents. You decide to use topic modeling. Identify two text-cleaning steps you would include in preprocessing of the text and briefly explain why they would be useful.

7.2 Content-Based Filtering

Briefly explain how content-based filtering works in the context of recommender systems.

7.3 Recommender Systems

Describe a scenario where collaborative-based filtering for a recommender system would have an advantage over content-based filtering.

8 Time Series

For help review Assignment 8 material.

8.1 Splitting the Data

You are working with CitiBike to train a model to forecast the number of bike rentals next month. CitiBike provided your team with a dataset containing rental history information formatted as a time-ordered sequence of data points. To evaluate the model's ability to generalize, your team divides the time series data into training and testing subsets as follows:

```
citibike = pd.read_csv("data/citibike.csv", index_col=0)
train_df, test_df = train_test_split(citibike, test_size=0.2, random_state=123)
```

What will happen when we split the data this way? Briefly explain.