0.1 Submission instructions

rubric={mechanics}

You receive marks for submitting your lab correctly, please follow these instructions:

Follow the general lab instructions.

Click here to view a description of the rubrics used to grade the questions

Make at least three commits.

Push your .ipynb file to your GitHub repository for this lab and upload it to Gradescope.

Before submitting, make sure you restart the kernel and rerun all cells.

Make sure to only make one gradescope submission per group, and to assign all group members on gradescope at submission time.

Also upload a .pdf export of the notebook to facilitate grading of manual questions (preferably WebPDF, you can select two files when uploading to gradescope)

Don't change any variable names that are given to you, don't move cells around, and don't include any code to install packages in the notebook.

The data you download for this lab SHOULD NOT BE PUSHED TO YOUR REPOSITORY (there is also a .gitignore in the repo to prevent this).

Include a clickable link to your GitHub repo for the lab just below this cell

It should look something like this https://github.ubc.ca/MDS-2020-21/DSCI_531_labX_yourcwl.

Points: 2

https://github.com/will-chh/573-lab4-creditcard-default-predictor#

0.2 1. Pick your problem and explain the prediction problem

rubric={reasoning}

In this mini project, you will pick one of the following problems:

1. A classification problem of predicting whether a credit card client will default or not. For this problem, you will use Default of Credit Card Clients Dataset. In this data set, there are 30,000 examples and 24 features, and the goal is to estimate whether a person will default (fail to pay) their credit card bills; this column is labeled "default.payment.next.month" in the data. The rest of the columns can be used as features. You may take some ideas and compare your results with the associated research paper, which is available through the UBC library.

OR

2. A regression problem of predicting `reviews_per_month`, as a proxy for the popularity of the listing with New York City Airbnb listings from 2019 dataset. Airbnb could use this sort of model to predict how popular future listings might be before they are posted, perhaps to help guide hosts create more appealing listings. In reality they might instead use something like vacancy rate or average rating as their target, but we do not have that available here.

Your tasks:

1. Spend some time understanding the problem and what each feature means. Write a few sentences on your initial thoughts on the problem and the dataset.
2. Download the dataset and read it as a pandas dataframe.
3. Carry out any preliminary preprocessing, if needed (e.g., changing feature names, handling of NaN values etc.)

Points: 3

Our problem of interest is predicting whether a credit card client will default on their payment next month. Given a dataset of 30,000 examples and 24 features, we aim to build a binary classification model that can accurately predict default payment behavior. This can be used by financial institutions in areas such as risk assessment and decision-making regarding credit issuance.

- ID: row identifier
- LIMIT_BAL: amount of given credit in NT dollars (includes individual and family/supplementary credit)
- SEX: gender (1=male, 2=female)
- EDUCATION: education level (1=graduate school; 2=university; 3=high school; 4=others)
- MARRIAGE: marital status (1=married; 2=single; 3=others)

- AGE: age in years

- PAY_0 to PAY_6: history of past monthly payment (from September 2005 to April 2005). Values are -1=pay duly (no delay), 1=payment delay for one month, 2=payment delay for two months, and so on.
- PAY_AMT1 to PAY_AMT6: amount of previous payment (payment status from the last 6 months starting from September)
- BILL_AMT1 to BILL_AMT6: amount of bill statement (from September 2005 to April 2005)
- default.payment.next.month: default payment (1=yes, 0=no). This is our target variable

There are a mix of categorical and numerical features in the dataset, demographic features include SEX, EDUCATION, MARRIAGE, and AGE. Financial features include LIMIT_BAL, PAY_0 to PAY_6, PAY_AMT1 to PAY_AMT6, and BILL_AMT1 to BILL_AMT6. Behavioural features could include payment history (PAY_0 to PAY_6) and previous payment amounts (PAY_AMT1 to PAY_AMT6). An initial thought is that PAY_0 to PAY_6 may be most predictive of default payment next month, as they likely reflect the client's payment behaviour over the past six months.

Package Importation

In [2]:
```
# Data & plotting
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import altair as alt
from ydata_profiling import ProfileReport

# Scikit-learn utilities
from sklearn.model_selection import (
    train_test_split,
    cross_val_score,
    cross_validate,
    GridSearchCV,
    RandomizedSearchCV,
)

# Preprocessing / transformations
from sklearn.preprocessing import OneHotEncoder, OrdinalEncoder, StandardScaler, KBinsDiscretiz
from sklearn.impute import SimpleImputer
from sklearn.compose import ColumnTransformer, make_column_transformer
from sklearn.pipeline import Pipeline, make_pipeline

# Models - classification & regression
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor, GradientBoostingCla
from sklearn.dummy import DummyClassifier, DummyRegressor

# Additional things to import
from scipy.stats import loguniform, randint
```

```
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.inspection import permutation_importance

from sklearn.metrics import (
    accuracy_score, precision_score, recall_score, f1_score, roc_auc_score,
    confusion_matrix, ConfusionMatrixDisplay
)
```

Muting Warnings

In [3]:
```
import warnings
warnings.filterwarnings(
    "ignore",
    message=r"You passed a .* to `is_pandas_dataframe`",
    module="altair.utils.data",
    category=UserWarning,
)
```

Data Loading

In [5]:
```
cr_card_df = pd.read_csv('data_file/UCI_Credit_Card.csv', header=0)
#cr_card_df = cr_card_df.sample(frac=0.25, random_state=123)
```

In [ ]:
```
#viewing the first 5 rows of data
cr_card_df.head()
```

In [ ]:
```
#checking for null values
cr_card_df.info() #checking datatypes and non-null counts
cr_card_df.isna().sum()
```

In [ ]:
```
#cleaning column names so that they are more consistent and readable (in snake_case)
cr_card_df.columns = cr_card_df.columns.str.lower().str.replace('.','_')
print(cr_card_df.columns)
```

In [ ]:
```
#dropping irrelevant columns
cr_card_df = cr_card_df.drop(columns=['id'])
```

In [ ]:
```
#according to the research paper, there are some extra categories that are not represented corr
print(cr_card_df['education'].value_counts().sort_index()) #5,6 are undefined/unknown, can be t
print(cr_card_df['marriage'].value_counts().sort_index()) #0 is undefined, can be treated as 'o
```

```
cr_card_df['education'] = cr_card_df['education'].replace({5:4, 6:4, 0:4})
cr_card_df['marriage'] = cr_card_df['marriage'].replace({0:3})
```

In [ ]:
```
#checking values after they are grouped to 'others' categories
print(cr_card_df['education'].value_counts().sort_index())
print(cr_card_df['marriage'].value_counts().sort_index())
```

In [ ]:
```
#checking class imbalance
cr_card_df['default_payment_next_month'].value_counts(normalize=True)
```

0.3 2. Data splitting

rubric={reasoning}

Your tasks:

1. Split the data into train and test portions.

   Make the decision on the `test_size` based on the capacity of your laptop.

Points: 1

```
In [ ]: # splitting into 70% train and 30% test portions
        cr_card_train, cr_card_test = train_test_split(
            cr_card_df,
            test_size=0.3,
            random_state=123
            )

        # Below could be dropped
        X_train = cr_card_train.drop(columns=['default_payment_next_month'])
        y_train = cr_card_train['default_payment_next_month']
        X_test = cr_card_test.drop(columns=['default_payment_next_month'])
        y_test = cr_card_test['default_payment_next_month']
```

0.4 3. EDA

rubric={viz,reasoning}

Perform exploratory data analysis on the train set.

Your tasks:

1. Include at least two summary statistics and two visualizations that you find useful, and accompany each one with a sentence explaining it.
2. Summarize your initial observations about the data.
3. Pick appropriate metric/metrics for assessment.

Points: 6

2. The profile report shows that there are no missing values in the dataset. The target variable, default_payment_next_month, has a class imbalance with 22.3% of instances being 1 (default) and 77.7% being 0 (no default). This indicates that the model might need to account for this imbalance during training.
   In the correlations section, we can see that pay_0 has the highest positive correlation with the target variable, suggesting that recent payment history is a strong predictor of default payment next month. BILL_AMT1 also shows a moderate positive correlation, indicating that higher bill amounts may be associated with a higher likelihood of default.
   Many financial features like bill_amt, pay_amt are highly skewed to the right, indicating that most clients have lower amounts while a few have very high amounts. This suggests that transformations like log transformation might be beneficial for these features.
3. Based on the class imbalance in the target variable, accuracy may not be sufficient to evaluate model performance. Therefore, we will use recall, which indicates how many actual defaults were correctly identified by the model, since missed defaults are the most expensive and important to identify. We can consider changing the decision threshold to optimize recall. Additionally, we will also report precision and F1-score to provide a more comprehensive evaluation of the model's performance (including a confusion matrix), as well as the ROC-AUC score to assess the model's ability to discriminate between classes across different thresholds.

```
In [ ]: profile = ProfileReport(cr_card_train, title='Credit card default training set EDA', explorative
        profile
```

```
In [ ]: cr_card_train.info()
```

```
In [ ]: #Creating a Correlation Table for all numerical features against our target column

        num_cols0 = cr_card_train.select_dtypes(include=['float64', 'int64']).columns
```

```
corr_w_target = (cr_card_train[num_cols0]
                 .corrwith(cr_card_train['default_payment_next_month'].astype(int))
                 .sort_values(ascending=False))

table = corr_w_target.reset_index()
table.columns = (['feature', 'correlation with default'])

table
```

Brief interpretation on Correlation outcome

1. With pay_0 at 0.3251, customers with worse repayment status last month (higher PAY_0) are more likely to default next month.
2. With education at 0.0326, age at 0.0107, and sex at -0.0463, these 3 fields do not clearly increase or decrease default risk, therefore, are considered to be dropped.
3. With limit_bal at -0.1492, customers with higher credit limit has lower chance of defaulting, which makes sense in real life.
4. All bill_amt* fields have small negative values, indicating a higher billed amount are associated with lower default risks.
5. All pay_amt* fields have negatives values, with pay_amt1 = -0.7156, it indicates higher and more recent payments are associated with lower default risks.

In [ ]:
```
#Creating a summary table of numeric features for each class (of our target)
num_feats = cr_card_train.select_dtypes(include=['float64']).columns

cr_card_train[['default_payment_next_month'] + list(num_feats)].melt(
    id_vars='default_payment_next_month',
    var_name='feature',
    value_name='value'
).pivot_table(
    index='feature', columns='default_payment_next_month',
    values='value',
    aggfunc=['mean', 'min', 'max', 'std'], observed=False
).round(2)
```

Brief interpretation on the summary table:

1. People who do NOT default pay more and have higher limits.
2. Defaulters have: lower credit limits, lower bill amounts, and lower historical payment amounts.

In [ ]:
```
#given that pay_* are our strongest features, visualizing the distribution of values within eac
alt.data_transformers.disable_max_rows()
```

```
pay_cols = ['pay_0', 'pay_2', 'pay_3', 'pay_4', 'pay_5', 'pay_6']

pay_long = cr_card_train[pay_cols + ['default_payment_next_month']].melt(
    id_vars='default_payment_next_month',
    var_name='pay_month',
    value_name='delay_status')

alt.Chart(pay_long).mark_bar().encode(
    x=alt.X('delay_status:O'),
    y=alt.Y('count():Q', title='Count of Clients'),
    color=alt.Color('default_payment_next_month:N', title='default')
).facet('pay_month:N', columns=3).properties(title='Payment Status distribution by month')
```

In [ ]:
```
categorical_feats = ['sex', 'education', 'marriage']

cat_data = cr_card_train[categorical_feats + ['default_payment_next_month']].melt(
    id_vars='default_payment_next_month',
    var_name='feature',
    value_name='value')

alt.Chart(cat_data).mark_bar().encode(
    x=alt.X('value:N'),
    y=alt.Y('count():Q', title='Count of Clients'),
    color=alt.Color('default_payment_next_month:N', title="default")
).facet('feature:N', columns=3).properties(title='Categorical feature distribution')
```

In [ ]:
```
#A deeper dive into the correlation between amount of bill statement and repayment status
num_feats = cr_card_train.select_dtypes(include=['float64']).columns

corr = cr_card_train[num_feats].corr()

corr_long = corr.reset_index().melt(
    id_vars='index',
    var_name='feature',
    value_name='correlation')

heatmap = alt.Chart(corr_long).mark_rect().encode(
    x=alt.X('feature:O', title='feature'),
    y=alt.Y('index:O', title='feature'),
    color=alt.Color('correlation:Q')
)

heatmap + heatmap.mark_text().encode(
    text=alt.Text('correlation:Q', format='.1f'),
    color=alt.value('black')
)
```

Brief Interpretation on the Correlation Heat Map Key Info only: 1. Bill_amt1 - 6 are highly correlated, leading to multicollinearity issue. 2. Pay_amt1 - 6 are somewhat correlated, multicollinearity is

still a concern. 3. Limit_bal has some correlation with other fields, it could be a very useful feature.

In [ ]:
```
import altair as alt

# Ensure default is categorical with readable labels
cr_card_train['default_label'] = cr_card_train['default_payment_next_month'].map({0: 'No Defaul

box = (
    alt.Chart(cr_card_train)
    .mark_boxplot(size=60)
    .encode(
        x=alt.X('default_label:N', title='Default Status'),
        y=alt.Y('limit_bal:Q', title='Credit Limit'),
        color=alt.Color('default_label:N', title='Default Status')
    )
    .properties(
        width=300,
        height=350,
        title='Boxplot of Credit Limit by Default Status'
    )
)

box
```

0.5 4. Feature engineering (Challenging)

rubric={reasoning}

Your tasks:

1. Carry out feature engineering. In other words, extract new features relevant for the problem and work with your new feature set in the following exercises. You may have to go back and forth between feature engineering and preprocessing.

Points: 0.5

Feature Engineering Reasoning: The following 3 new features are created as a sniff test:

1. payment difference: pay_0 - pay_6 captures the change in payment behaviour from 6months ago to this month. A negative value indicates improvement (since pay_6 was bad and pay_0 was good), while a positive value indicates worsening behaviour, which may be predictive of default risk (behavioural trend over the 6 mo)
2. average pay amount: avg_pay_amt = captures the average payment amount over the past 6 months, which may indicate the client's ability to pay the next one
3. standard deviation of pay: captures payment volatility over the past 6 months, which may indicate inconsistent payment behaviour (for example: high standard deviation 0,1,3,5,0,6 -> erratic payment behaviour -> higher default risk). Low std (1,1,1,1,1,1) indicates consistent payment behaviour -> lower default risk

In [ ]:
```
#aggregate features for pay*

    #Creating copies on the cr_card_train before having new features
    cr_card_train_fe = cr_card_train.copy()
    cr_card_test_fe = cr_card_test.copy()

    #updating this line to have pay_0 as its own feature as pay_0 deviates from the rest of the pay
    #averaging only from pay_2 to pay_6
    hist_pay_cols = ['pay_2', 'pay_3', 'pay_4', 'pay_5', 'pay_6']

    # Train
    cr_card_train_fe['pay_avg_2_6'] = cr_card_train_fe[hist_pay_cols].mean(axis=1)
    cr_card_train_fe['pay_past_std'] = cr_card_train_fe[hist_pay_cols].std(axis=1)
    cr_card_train_fe['pay_diff'] = cr_card_train_fe['pay_0'] - cr_card_train_fe['pay_6']

    # Test
    cr_card_test_fe['pay_avg_2_6'] = cr_card_test_fe[hist_pay_cols].mean(axis=1)
    cr_card_test_fe['pay_past_std'] = cr_card_test_fe[hist_pay_cols].std(axis=1)
```

```
cr_card_test_fe['pay_diff'] = cr_card_test_fe['pay_0'] - cr_card_test_fe['pay_6']


cr_card_train_fe.head()
```

In [ ]:
```
eng_cols = ['pay_avg_2_6', 'pay_past_std', 'pay_diff']

eng_long = cr_card_train_fe[eng_cols + ['default_payment_next_month']].melt(
    id_vars='default_payment_next_month',
    var_name='feature',
    value_name='value')

alt.Chart(eng_long).mark_boxplot(size=40).encode(
    x=alt.X('default_payment_next_month:O'),
    y=alt.Y('value:Q'),
    color=alt.Color('default_payment_next_month:N'),
    facet=alt.Facet('feature:N', columns=3)
).properties(width=180, height=240, title='Distribution of engineered payment features')
```

In [ ]:
```
eng_long = cr_card_train_fe[eng_cols + ['default_payment_next_month']].melt(
    id_vars='default_payment_next_month',
    var_name='feature',
    value_name='value')

alt.Chart(eng_long).mark_bar(opacity=0.8).encode(
    x=alt.X('value:Q', bin=alt.Bin(maxbins=30)),
    y=alt.Y('count():Q'),
    color=alt.Color('default_payment_next_month:N'),
    facet=alt.Facet('feature:N', columns=3)
).properties(width=180, height=240, title='Distribution of engineered payment features')
```

0.6 5. Preprocessing and transformations

rubric={accuracy,reasoning}

Your tasks:

1. Identify different feature types and the transformations you would apply on each feature type.
2. Define a column transformer, if necessary.

Points: 4

Preprocessing Set up In Q4, it explicitly mentioned to work with the newly created features in the following exercises, therefore we will be using using the transformed data sets for the rest of the assignment.

In [ ]:
```
cr_card_test_fe.head()
```

In [ ]:
```
#Creating Copies here, all transformations will be working on datasets ending without *_fe
cr_card_train = cr_card_train_fe.copy()
cr_card_test = cr_card_test_fe.copy()


X_train = cr_card_train.drop(columns=['default_payment_next_month'])
y_train = cr_card_train['default_payment_next_month']

X_test = cr_card_test.drop(columns=['default_payment_next_month'])
y_test = cr_card_test['default_payment_next_month']


#Grouping feature types
numeric_feats = ['limit_bal',
                 'bill_amt1', 'bill_amt2', 'bill_amt3', 'bill_amt4', 'bill_amt5', 'bill_amt6',
                 'pay_amt1', 'pay_amt2', 'pay_amt3', 'pay_amt4', 'pay_amt5', 'pay_amt6',
                 'pay_avg_2_6', 'pay_past_std', 'pay_diff', # engineered
                 'age']

ordinal_feats = ['education', 'pay_0']

categorical_feats = ['sex', 'marriage']

drop_feats = ['pay_2', 'pay_3', 'pay_4', 'pay_5', 'pay_6']
```

In [ ]:
```
X_train = X_train.drop(columns=drop_feats)
X_test = X_test.drop(columns=drop_feats)
```

In [ ]:
```
#Applying transformations
preprocessor = make_column_transformer(
    (StandardScaler(), numeric_feats),
    (OrdinalEncoder(handle_unknown='use_encoded_value', unknown_value=-1), ordinal_feats),
    (OneHotEncoder(handle_unknown='ignore'), categorical_feats),
    #("drop", drop_feats)
)

#Fitting preprocessor on X_train only
preprocessor.fit(X_train)

#Transforming both data sets
X_train_enc = preprocessor.transform(X_train)
X_test_enc = preprocessor.transform(X_test)

#Collecting column names, ordering matters, pls refer above, numeric to ordinal to onehot
all_columns = preprocessor.get_feature_names_out()

X_train_df = pd.DataFrame(X_train_enc, columns=all_columns, index=X_train.index)
X_test_df  = pd.DataFrame(X_test_enc, columns=all_columns, index=X_test.index)

X_train_df
```

0.7 6. Baseline model

rubric={accuracy}

Your tasks: 1. Train a baseline model for your task and report its performance.

Points: 2

In [ ]:
```
import warnings
from sklearn.exceptions import UndefinedMetricWarning

warnings.filterwarnings("ignore", category=UndefinedMetricWarning)
```

In [ ]:
```
cross_val_results = {}

dummy = DummyClassifier(strategy = "most_frequent")
scoring_metrics = ['accuracy', 'f1', 'precision', 'recall', 'roc_auc']

cross_val_results['dummy'] = (pd.DataFrame(
    cross_validate(
        dummy,
        X_train_enc,
        y_train,
        scoring = scoring_metrics,
        cv=5,
        return_train_score=True))
    .agg(['mean', 'std'])
    .round(3)
    .T)

cross_val_results['dummy']
```

Brief Interpretation Accuracy score of 0.777 confirms the strong class imbalance.

0.8 7. Linear models

rubric={accuracy,reasoning}

Your tasks:

1. Try a linear model as a first real attempt.
2. Carry out hyperparameter tuning to explore different values for the regularization hyperparameter.
3. Report cross-validation scores along with standard deviation.
4. Summarize your results.

Points: 8

Linear Models Results Summary - Updated A Logistic Regression model with class_weight='balanced' was used as the first linear model for predicting credit card defaults. Hyperparameter tuning was performed using RandomizedSearchCV to optimize the regularization strength, and the best value of C was approximately 0.015. This relatively small value indicates that stronger regularization helped stabilize the model and prevent overfitting. The tuned model achieved a cross-validated test accuracy of about 0.721 with a low standard deviation across folds, suggesting that performance was stable and not sensitive to the specific subset of training data used. More importantly, the tuned model achieved an F1 score of approximately 0.506, with precision around 0.418 and recall around 0.642. Because recall is especially important in identifying customers who will default, the tuned model's ability to capture roughly 64% of the actual defaulters represents a substantial improvement over the Dummy Classifier baseline. The ROC AUC of 0.735 further indicates that the model has reasonable discriminative ability between default and non-default cases.

Comparing the tuned model to the untuned version reveals relatively small but meaningful improvements. The untuned model, which uses the default C = 1.0, produced a slightly weaker cross-validation score and showed marginally poorer calibration when predicting the positive class. The confusion matrices illustrate this difference clearly: the tuned model reduced the number of false positives from 1,871 to 1,860 while increasing the number of true negatives by the same amount. True positives and false negatives remained unchanged across both models, meaning the tuned regularization primarily adjusted the model's confidence on borderline negative-class predictions. Although these changes appear small, they are consistent with the tuning objective and reflect the limited flexibility of a linear model on this dataset. Overall, the tuned logistic regression model provides stable performance, improves recall compared to the untuned version, and establishes a reasonable, well-regularized linear baseline for subsequent comparison with more flexible nonlinear models.

In [ ]:
```
#Model 1 - Logistics Regression

        pipe_lr = make_pipeline(
            preprocessor,
            LogisticRegression(class_weight='balanced', max_iter=1000)
            )

        # Hyperparameter random search space for C (inverse regularization strength), on a log scale
```

```
param_dist = {"logisticregression__C": loguniform(0.01, 10)}

random_search = RandomizedSearchCV(
    pipe_lr,
    param_dist,
    n_iter=20,
    cv=5,
    n_jobs=-1,
    random_state=123,
    return_train_score=True
    )

# Carrying out random search, raw X_train is fine
random_search.fit(X_train, y_train)

# Extract the best pipeline (preprocessor + tuned LogisticRegression)
pipe_lr_tuned = random_search.best_estimator_

logreg_cv_scores = pd.DataFrame(cross_validate(
    pipe_lr_tuned,
    X_train, y_train,
    cv=5,
    scoring = scoring_metrics,
    return_train_score=True
)).agg(['mean', 'std']).round(3).T

cross_val_results['logreg'] = logreg_cv_scores

cross_val_results['logreg']
```

In [ ]:
```
print("Best Parameters:", random_search.best_params_)
```

In [ ]:
```
print("CV Score:", random_search.best_score_)
```

In [ ]:
```
# Untuned Logistics Regression Pipeline
pipe_lr.fit(X_train, y_train)

predictions = pipe_lr.predict(X_test)

confusion_matrix(y_test, predictions)
```

In [ ]:
```
# Tuned Logistics Regression Pipeline
pipe_lr_tuned.fit(X_train, y_train)

predictions = pipe_lr_tuned.predict(X_test)

confusion_matrix(y_test, predictions)
```

In [ ]:
```
%matplotlib inline
```

In [ ]:
```
from sklearn.metrics import ConfusionMatrixDisplay, confusion_matrix

# Plot confusion matrix on the TUNED Model, there is barely any improvement.
disp = ConfusionMatrixDisplay.from_estimator(
    pipe_lr_tuned,
    X_test,
    y_test,
    display_labels=["No Default", "Default"],
    cmap="Blues",
    values_format="d",
)

disp.ax_.set_title("Logistic Regression - Confusion Matrix (Test Set)")
plt.show()
```

0.9 8. Different models

rubric={accuracy,reasoning}

Your tasks: 1. Try out three other models aside from the linear model. 2. Summarize your results in terms of overfitting/underfitting and fit and score times. Can you beat the performance of the linear model?

Points: 10

Interpretation of Non-Linear Models (SVC, Gradient Boosting, Random Forest) To compare non-linear models against the tuned logistic regression baseline, three additional classifiers were evaluated: an SVC with an RBF kernel, a Gradient Boosting Classifier, and a Random Forest with class balancing. All models were evaluated using identical 5-fold cross-validation and the same preprocessing pipeline, allowing for a fair comparison across accuracy, F1, precision, recall, ROC AUC, and fit/score times.

Across the board, the non-linear models achieved higher accuracy and ROC AUC than logistic regression. Gradient Boosting obtained the highest test accuracy (0.821), with Random Forest close behind (0.815). SVC also outperformed logistic regression in both accuracy (0.767 vs. 0.721) and ROC AUC (0.759 vs. 0.735), confirming that non-linear decision boundaries better fit the complex structure of the data than a linear model. However, when focusing on metrics relevant to the minority (default) class—particularly recall and F1—the story becomes more nuanced. Logistic regression continues to outperform the non-linear models in recall (0.642 for LR vs. 0.607 for SVC, 0.377 for GBC, and 0.352 for RF). Even though the tree-based models achieve higher precision (0.66–0.71), they identify substantially fewer actual defaults, resulting in lower F1 scores. This reflects a fundamental tradeoff: boosting and random forests tend to be conservative in predicting the minority class unless explicitly tuned for recall.

From an overfitting perspective, the Random Forest shows signs of severe overfitting: training accuracy, F1, precision, recall, and ROC AUC are all effectively 1.000, while test scores drop sharply (e.g., recall falling to 0.352). This indicates that the forest memorizes the training data but struggles to generalize to unseen examples. Gradient Boosting shows milder overfitting, with training accuracy (0.829) slightly exceeding test accuracy (0.821), and similarly small but consistent gaps across other metrics. SVC falls in between—its gap between train and test scores is modest but noticeable, consistent with the flexible non-linear decision boundary of the RBF kernel. In contrast, logistic regression shows almost no gap between train and test performance, indicating a very stable and well-regularized model.

Fit and score time differences also help explain model behavior. SVC is by far the slowest model, with average fit times near 11 seconds per fold and score times around 7 seconds, reflecting the computational cost of RBF kernel distance calculations on roughly 30,000 samples. Gradient Boosting is faster (5.3 s fit), while Random Forest is significantly faster still (2.9 s fit). Logistic regression remains extremely efficient at 0.038 seconds, making it attractive when computational resources or latency matter.

One surprising observation is that the tree-based models—particularly Gradient Boosting and Random Forest—have much worse recall than logistic regression, even though they often outperform LR on accuracy and ROC AUC. This occurs because tree models trained with default hyperparameters tend to prioritize splits that maximize accuracy rather than detect minority-class instances. Even with class_weight='balanced', Random Forest tends to make conservative positive predictions, resulting in high precision but low recall. Logistic regression, by contrast, distributes probability mass more evenly between classes and, combined

with class balancing, becomes better at detecting default cases. With hyperparameter tuning (e.g., adjusting decision thresholds, increasing the number of boosted trees, reducing RF depth), tree models can be pushed toward better recall, but with default settings they tend to underpredict the minority class.

Summary Non-linear models outperform logistic regression in accuracy and ROC AUC, but logistic regression remains superior on the recall of the default class, which is the most important business metric in this domain. Random Forest overfits heavily, Gradient Boosting overfits mildly, and SVC shows moderate overfitting but reasonably strong balanced performance. Logistic Regression remains the most stable and interpretable model, even if not the most accurate, demonstrating why model evaluation must go beyond accuracy alone when dealing with imbalanced classification problems such as credit default prediction.

In [ ]:
```
# First Model: SVC Classifier

pipe_svc = make_pipeline(
    preprocessor,
    SVC(class_weight='balanced')
    )

cross_val_results['SVC'] = pd.DataFrame(cross_validate(
    pipe_svc,
    X_train,
    y_train,
    return_train_score=True,
    scoring = scoring_metrics,
    cv = 5
    )).agg(['mean', 'std']).round(3).T

cross_val_results['SVC']
```

In [ ]:
```
# Second Model: Gradient Boosting Classifier

pipe_GBC = make_pipeline(
    preprocessor,
    GradientBoostingClassifier(random_state=123)
    )

cross_val_results['GBC'] = pd.DataFrame(cross_validate(
    pipe_GBC,
    X_train,
    y_train,
    return_train_score=True,
    scoring = scoring_metrics,
    cv = 5
    )).agg(['mean', 'std']).round(3).T

cross_val_results['GBC']
```

In [ ]:
```
# Third Model: Random Forest Classifier
pipe_tree = make_pipeline(
```

```
        preprocessor,
        RandomForestClassifier(class_weight="balanced", random_state=123)
        )

    cross_val_results['RF'] = pd.DataFrame(cross_validate(
        pipe_tree,
        X_train,
        y_train,
        return_train_score=True,
        scoring = scoring_metrics,
        cv = 5
        )).agg(['mean', 'std']).round(3).T

    cross_val_results['RF']
```

In [ ]:
```
cross_val_results_df = pd.concat(cross_val_results, axis=1)
cross_val_results_df
```

0.10 9. Feature selection (Challenging)

rubric={reasoning}

Your tasks:

Make some attempts to select relevant features. You may try `RFECV`, forward/backward selection or L1 regularization for this. Do the results improve with feature selection? Summarize your results. If you see improvements in the results, keep feature selection in your pipeline. If not, you may abandon it in the next exercises unless you think there are other benefits with using fewer features.

Points: 0.5

Type your answer here, replacing this text.

In [ ]: ...

In [ ]: ...

In [ ]: ...

0.11 10. Hyperparameter optimization

rubric={accuracy,reasoning}

Your tasks:

Make some attempts to optimize hyperparameters for the models you've tried and summarize your results. In at least one case you should be optimizing multiple hyperparameters for a single model. You may use `sklearn`'s methods for hyperparameter optimization or fancier Bayesian optimization methods. Briefly summarize your results. - GridSearchCV
- RandomizedSearchCV - scikit-optimize

Points: 6

Summary of Tuned Model Results Tuning the three non-linear models led to modest but meaningful shifts in performance, with the impact varying by model family. SVC tuning primarily adjusted the regularization (C) and kernel smoothness (gamma), but the overall performance remained almost identical to the untuned version. Test accuracy and ROC AUC stayed essentially unchanged, and the recall improved only slightly (0.607 → 0.609). This reflects the fact that the baseline SVC model was already close to its optimal operating point, and the limited gains came at the cost of dramatically increased fit time (from ~11 seconds to ~58 seconds per fold).

For Gradient Boosting, tuning learning rate and depth similarly produced small refinements rather than dramatic improvements. Test accuracy remained 0.821 before and after tuning, and ROC AUC increased only marginally (0.780 → 0.781). Precision remained high (~0.68), but recall continued to lag the linear model at about 0.38. These results suggest that the default hyperparameters of gradient boosting were already well-calibrated to the dataset, and recall performance is limited more by the model's conservative bias toward predicting the minority class than by its depth or learning rate.

Random Forest tuning produced the most noticeable changes. The tuned model substantially reduced overfitting: train accuracy dropped from an unrealistic 0.999 to a more reasonable 0.811, and train ROC AUC fell from 1.000 to 0.859. This indicates that tuning depth and leaf size successfully constrained the model. As a result, test performance improved across all key metrics. Recall increased from 0.352 to 0.593, precision improved slightly, and ROC AUC rose from 0.761 to 0.782, now matching or slightly surpassing Gradient Boosting. Although accuracy decreased slightly (0.815 → 0.784), the tuned model exhibits far more balanced generalization and a significantly stronger ability to identify defaulting customers.

Overall, tuning did not fundamentally reorder the models' performance rankings, but it did produce important refinements—especially for Random Forest, where tuning meaningfully reduced overfitting and achieved substantial improvements in recall and ROC AUC. These tuned models provide a more reliable foundation for interpretation and test-set evaluation in later questions.

In [ ]:
```
#SVC tune C and gamma

        # Pipeline: preprocess → SVC (RBF)
        pipe_svc = make_pipeline(
```

```
    preprocessor,
    SVC(class_weight='balanced', probability=True)  # probability=True needed for ROC AUC
)

# Search space for SVC
# C: controls margin softness (bigger C -> more complex model)
# gamma: controls RBF kernel width (bigger gamma -> more wiggly boundary)
param_dist_svc = {
    "svc__C": loguniform(1e-3, 1e2),
    "svc__gamma": loguniform(1e-4, 1e0),
}

# Randomized search over SVC hyperparameters
svc_search = RandomizedSearchCV(
    pipe_svc,
    param_distributions=param_dist_svc,
    n_iter=20,              # number of random combinations
    cv=5,
    scoring="roc_auc",      # optimize for ROC AUC
    n_jobs=-1,
    random_state=123,
    return_train_score=True,
)

svc_search.fit(X_train, y_train)

print("Best SVC params:", svc_search.best_params_)
print("Best SVC CV ROC AUC:", svc_search.best_score_)

pipe_svc_tuned = svc_search.best_estimator_
```

In [ ]:
```
#GBC tune learning_rate, max_depth, n_estimators

# Pipeline: preprocess → GradientBoosting
pipe_gbc = make_pipeline(
    preprocessor,
    GradientBoostingClassifier(random_state=123)
)

# Small grid search (tree depth, number of trees, learning rate)
param_grid_gbc = {
    "gradientboostingclassifier__learning_rate": [0.01, 0.05, 0.1],
    "gradientboostingclassifier__max_depth": [2, 3, 4],
    "gradientboostingclassifier__n_estimators": [100, 200],
}

gbc_search = GridSearchCV(
    pipe_gbc,
    param_grid=param_grid_gbc,
    cv=5,
    scoring="roc_auc",
    n_jobs=-1,
    return_train_score=True,
```

```
)

gbc_search.fit(X_train, y_train)

print("Best GBC params:", gbc_search.best_params_)
print("Best GBC CV ROC AUC:", gbc_search.best_score_)

pipe_gbc_tuned = gbc_search.best_estimator_
```

In [ ]:
```
#RF tune n_estimators, max_depth, min_samples_leaf

# Pipeline: preprocess → RandomForest
pipe_rf = make_pipeline(
    preprocessor,
    RandomForestClassifier(class_weight="balanced", random_state=123)
)

# RF search space
# n_estimators: number of trees
# max_depth: limit depth to reduce overfitting
# min_samples_leaf: make leaves less pure, improves generalization
param_dist_rf = {
    "randomforestclassifier__n_estimators": randint(200, 800),
    "randomforestclassifier__max_depth": [None, 5, 10, 20],
    "randomforestclassifier__min_samples_leaf": randint(1, 20),
}

rf_search = RandomizedSearchCV(
    pipe_rf,
    param_distributions=param_dist_rf,
    n_iter=20,
    cv=5,
    scoring="roc_auc",
    n_jobs=-1,
    random_state=123,
    return_train_score=True,
)

rf_search.fit(X_train, y_train)

print("Best RF params:", rf_search.best_params_)
print("Best RF CV ROC AUC:", rf_search.best_score_)

pipe_rf_tuned = rf_search.best_estimator_
```

In [ ]:
```
# Adding SVC tuned score

svc_cv_scores = (
    pd.DataFrame(
        cross_validate(
            pipe_svc_tuned,
```

```
                X_train,
                y_train,
                cv=5,
                scoring=scoring_metrics,
                return_train_score=True
            )
        )
        .agg(['mean', 'std'])
        .round(3)
        .T
    )

    cross_val_results['svc_tuned'] = svc_cv_scores
```

In [ ]:
```
# Adding GBC tuned score

gbc_cv_scores = (
    pd.DataFrame(
        cross_validate(
            pipe_gbc_tuned,
            X_train,
            y_train,
            cv=5,
            scoring=scoring_metrics,
            return_train_score=True
        )
    )
    .agg(['mean', 'std'])
    .round(3)
    .T
)

cross_val_results['gbc_tuned'] = gbc_cv_scores
```

In [ ]:
```
# Adding RF tuned score

rf_cv_scores = (
    pd.DataFrame(
        cross_validate(
            pipe_rf_tuned,
            X_train,
            y_train,
            cv=5,
            scoring=scoring_metrics,
            return_train_score=True
        )
    )
    .agg(['mean', 'std'])
    .round(3)
    .T
)
```

```
        cross_val_results['rf_tuned'] = rf_cv_scores
```

In [ ]:
```
cross_val_results_df = pd.concat(cross_val_results, axis=1)
cross_val_results_df
```

0.12 11. Interpretation and feature importances

rubric={accuracy,reasoning}

Your tasks:

1. Use the methods we saw in class (e.g., `permutation_importance` or `shap`) (or any other methods of your choice) to examine the most important features of one of the non-linear models.
2. Summarize your observations.

Points: 8

Why ROC AUC is best for comparing your non-linear models It does not depend on the probability threshold

It evaluates ranking ability, not class calibration

It handles imbalanced datasets gracefully

It is robust across different model families

It reflects general predictive power better than recall/precision alone

It is the metric emphasized in your lecture notes for fair model comparison

Permutation Importance on the tuned Random Forest Interpretation

The permutation importance results show that recent repayment behavior is by far the strongest signal for predicting credit default. PAY_0, the most recent delinquency indicator, is the top feature by a large margin, confirming that short-term repayment status is the most influential driver of model predictions. The engineered feature pay_avg_2_6, which summarizes repayment behavior across the previous five months, is the second-most important feature, indicating that longer-term repayment patterns also contribute substantial predictive power.

Credit limit (limit_bal) ranks next, suggesting that customers with lower credit limits may be more prone to default. Several payment amount variables (pay_amt1, pay_amt2, pay_amt3, etc.) and bill amounts (bill_amt1–bill_amt4, bill_amt6) appear prominently in the top 15, reflecting that both repayment capacity and outstanding balances influence default risk. The engineered volatility feature pay_diff and the standard deviation summary pay_past_std also carry meaningful importance, reinforcing that your feature engineering successfully captured useful behavioral trends.

Overall, the model relies primarily on recent and historical repayment behavior, secondarily on balance and payment amount patterns, and far less on demographic variables — a pattern consistent with typical credit-risk models.

In [ ]:
```
from sklearn.inspection import permutation_importance

rf_tuned = pipe_rf_tuned  # tuned RF pipeline

#Permutation importance on the pipeline
result = permutation_importance(
    rf_tuned,
    X_train,
    y_train,
    n_repeats=10,
    scoring='roc_auc',
    random_state=123
)

#Use original train column names
importances_mean = result.importances_mean
perm_sorted_idx = importances_mean.argsort()

perm_df = pd.DataFrame(
    {
        "feature": X_train.columns,
        "importance_mean": importances_mean
    }
).sort_values("importance_mean", ascending=False)

perm_df.head(15)
```

In [ ]:
```
# Plotting out bar chart of the top features
plt.figure(figsize=(8, 6))
plt.boxplot(
    result.importances[perm_sorted_idx].T,
    vert=False,
    labels=X_train.columns[perm_sorted_idx],
)
plt.xlabel("Permutation feature importance (Δ ROC AUC)")
plt.title("Tuned Random Forest - Permutation Importances (Train)")
plt.tight_layout()
plt.show()
```

0.13 12. Results on the test set

rubric={accuracy,reasoning}

Your tasks:

1. Try your best performing model on the test data and report test scores.
2. Do the test scores agree with the validation scores from before? To what extent do you trust your results? Do you think you've had issues with optimization bias?
3. Take one or two test predictions and explain them with SHAP force plots.

Points: 6

Interpretation of Test Scores The test-set results for the tuned Random Forest model closely match the validation scores obtained during cross-validation. Earlier, the model achieved a validation ROC AUC of approximately 0.782, and on the test set it reached 0.781, showing almost no degradation in performance. Accuracy (0.776), recall (0.593), and F1 score (0.534) are also consistent with the cross-validation metrics. This alignment suggests that the model generalizes well to unseen data and that we are not suffering from meaningful overfitting.

Because the validation and test scores are so similar, I have high confidence in these results. The hyperparameter tuning was performed using a separate training-only cross-validation loop, and the test set was only used once at the end, so the risk of optimization bias is low. The consistency between validation and test ROC AUC—our most reliable metric for imbalanced problems—reinforces that the tuning process produced a stable model rather than one overfitted to the training folds.

Interpretation of the SHAP values To understand why the model makes its predictions, I examined SHAP contributions for individual test-set customers. For a specific default case, the SHAP values indicate which features push the model toward predicting default (positive values) and which features push it toward predicting non-default (negative values).

For the example below, the model predicted default, and the SHAP values show that the strongest positive contributor was PAY_0, with a SHAP value of +0.194, meaning the customer's most recent repayment status strongly increased the model's belief that this person would default. Additional positive contributions came from pay_diff (difference between recent and past repayment severity) and pay_past_std, both suggesting inconsistent or deteriorating payment behavior.

In contrast, several features pulled the prediction downward (toward non-default). Higher credit limit, along with relatively stronger historical payments such as pay_avg_2_6, pay_amt3, and multiple bill amounts, all had negative SHAP values. These characteristics suggested financial stability, partially offsetting the risk signals but not enough to override the strong effect from PAY_0 and related repayment-history variables.

Overall, the SHAP force and waterfall plots reveal that the model's decisions are driven primarily by repayment behavior—especially the most recent month—while monetary features such as limit and payment

amounts provide secondary stabilizing effects. This aligns with the earlier permutation-importance results and helps confirm that the model is behaving in a financially intuitive way.

In [ ]:
```
# Predict on test set
y_pred = pipe_rf_tuned.predict(X_test)
y_proba = pipe_rf_tuned.predict_proba(X_test)[:, 1]

# Compute test metrics
test_metrics = {
    "accuracy": accuracy_score(y_test, y_pred),
    "precision": precision_score(y_test, y_pred),
    "recall": recall_score(y_test, y_pred),
    "f1_score": f1_score(y_test, y_pred),
    "roc_auc": roc_auc_score(y_test, y_proba)
}

pd.DataFrame(test_metrics, index=["RandomForest_Tuned"])
```

In [ ]:
```
#Plotting the confusion matrix here

disp = ConfusionMatrixDisplay.from_estimator(
    pipe_rf_tuned,
    X_test,
    y_test,
    display_labels=["No Default", "Default"],
    cmap="Blues",
    values_format="d"
)

disp.ax_.set_title("Confusion Matrix - Tuned Random Forest (Test Set)")
plt.show()
```

SHAP Force Plot

In [ ]:
```
# Get preprocessor and RF model from the tuned pipeline
ct = pipe_rf_tuned.named_steps["columntransformer"]
rf_model = pipe_rf_tuned.named_steps["randomforestclassifier"]

# Encode the test data (like X_test_enc in lecture)
X_test_enc = ct.transform(X_test)
feature_names = ct.get_feature_names_out()

# Reset y_test index to align with X_test_enc row indices
y_test_reset = y_test.reset_index(drop=True)

y_test_reset.head()
```

In [ ]:
```
# Indices for each class
```

```
no_default_idx = y_test_reset[y_test_reset == 0].index.tolist()
default_idx    = y_test_reset[y_test_reset == 1].index.tolist()

# Pick one example of each (10th in each list just like lecture)
ex_no_default_index = no_default_idx[10]
ex_default_index    = default_idx[10]
```

In [ ]:
```
import shap

# TreeExplainer on the RF model
rf_explainer = shap.TreeExplainer(rf_model)

# Explanation object for all test rows (in encoded space)
rf_explanation = rf_explainer(X_test_enc)   # shape: (n_samples, n_features)
```

In [ ]:
```
rf_explanation
```

In [ ]:
```
# SHAP values for one example (e.g., a default case)
ex_idx = ex_default_index  # or ex_no_default_index

shap_vals_ex = rf_explanation[ex_idx].values[:, 1]

pd.DataFrame(
    shap_vals_ex,
    index=feature_names,
    columns=["SHAP values"],
).sort_values("SHAP values")
```

In [ ]:
```
ex_idx = ex_default_index

# Base value for this example & class 1 (default)
base_val = rf_explanation.base_values[ex_idx, 1]

# SHAP values for this example & class 1
shap_vals = rf_explanation.values[ex_idx, :, 1]

# Force plot (no features argument needed)
shap.plots.force(base_val, shap_vals, matplotlib=True)
```

0.14 13. Summary of results

rubric={reasoning}

Imagine that you want to present the summary of these results to your boss and co-workers.

Your tasks:

1. Create a table summarizing important results.
2. Write concluding remarks.
3. Discuss other ideas that you did not try but could potentially improve the performance/interpretability .
4. Report your final test score along with the metric you used at the top of this notebook.

Points: 8

Concluding Remarks Across all experiments, the tuned Random Forest achieved the strongest balance of predictive power and generalization. Its test ROC AUC of 0.781 closely matches its validation performance, indicating reliable generalization and low risk of overfitting. Recall (0.593) and F1 (0.534) demonstrate that the model captures a meaningful proportion of true defaults while maintaining reasonable precision. SHAP and permutation-importance analyses both highlight repayment behavior—especially PAY_0 and recent payment patterns—as the dominant predictors of default risk, which is consistent with financial intuition.

Overall, the model provides actionable predictive performance while maintaining transparency into the main risk drivers.

Future Improvements Several enhancements could further improve model performance or interpretability:

1. Threshold tuning: Adjusting the classification threshold could improve recall or precision depending on business requirements (e.g., reducing missed defaults).
2. Cost-sensitive learning: Incorporating asymmetric costs of false negatives vs false positives may better align predictions with real financial risk.
3. Feature interactions: Tree-based models implicitly capture interactions, but explicitly engineering domain-specific interactions could strengthen signal.
4. Alternative models: XGBoost or LightGBM often outperform Random Forests on tabular data and support built-in handling of class imbalance.
5. More granular temporal features: Breaking down repayment history into slope/trend features may capture behavioral changes more effectively.

These directions could be explored in future iterations if higher recall or tighter risk ranking is desired.

0.15 Summary Table

```
In [ ]: # Select only test metrics from your cross_val_results_df
        metrics_to_keep = [
            "test_accuracy",
            "test_precision",
            "test_recall",
            "test_f1",
            "test_roc_auc"
        ]

        summary_table = cross_val_results_df.loc[metrics_to_keep].xs("mean", level=1, axis=1)

        # Pretty formatting
        summary_table = summary_table.T.round(3)
        summary_table
```

0.16 Final Results

```
In [ ]: final_results = {
            "accuracy": 0.776111,
            "precision": 0.485907,
            "recall": 0.592916,
            "f1_score": 0.534104,
            "roc_auc": 0.781173
        }

        final_results_df = pd.DataFrame(final_results, index=["RandomForest_Tuned"]).T
        final_results_df
```

0.17 14. Creating a data analysis pipeline (Challenging)

rubric={reasoning}

Your tasks:

- Convert this notebook into scripts to create a reproducible data analysis pipeline with appropriate documentation. Submit your project folder in addition to this notebook on GitHub and briefly comment on your organization in the text box below.

Points: 0.5

Type your answer here, replacing this text.

0.18 15. Your takeaway from the course (Challenging)

rubric={reasoning}

Your tasks:

What is your biggest takeaway from this course?

Points: 0.5

The biggest takeaway from this course is that there are various ways to improve model performance. One key aspect is selecting the appropriate evaluation metrics, recognizing that accuracy is not always the best measure, and that metrics like precision and recall can provide more context and meaningful insights depending on the problem. Another important approach is feature engineering, which plays a critical role in building effective models and often requires creativity from the data scientist to design new features that can meaningfully improve performance.