Credit Risk Resampling Techniques

In [1]:

```
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```
import numpy as np
import pandas as pd
from pathlib import Path
from collections import Counter
```

Read the CSV and Perform Basic Data Cleaning

In [3]:

```
columns = [
    "loan_amnt", "int_rate", "installment", "home_ownership",
    "annual_inc", "verification_status", "issue_d", "loan_status",
    "pymnt_plan", "dti", "delinq_2yrs", "inq_last_6mths",
    "open_acc", "pub_rec", "revol_bal", "total_acc",
    "initial_list_status", "out_prncp", "out_prncp_inv", "total_pymnt",
    "total_pymnt_inv", "total_rec_prncp", "total_rec_int", "total_rec_late_fe
    "recoveries", "collection_recovery_fee", "last_pymnt_amnt", "next_pymnt_d
    "collections_12_mths_ex_med", "policy_code", "application_type", "acc_now
    "tot_coll_amt", "tot_cur_bal", "open_acc_6m", "open_act_il",
    "open_il_12m", "open_il_24m", "mths_since_rcnt_il", "total_bal_il",
    "il_util", "open_rv_12m", "open_rv_24m", "max_bal_bc",
    "all_util", "total_rev_hi_lim", "inq_fi", "total_cu_tl",
    "inq_last_12m", "acc_open_past_24mths", "avg_cur_bal", "bc_open_to_buy",
    "bc_util", "chargeoff_within_12_mths", "delinq_amnt", "mo_sin_old_il_acct
    "mo_sin_old_rev_tl_op", "mo_sin_rcnt_rev_tl_op", "mo_sin_rcnt_tl", "mort_
    "mths_since_recent_bc", "mths_since_recent_inq", "num_accts_ever_120_pd",
    "num_actv_rev_tl", "num_bc_sats", "num_bc_tl", "num_il_tl",
    "num_op_rev_tl", "num_rev_accts", "num_rev_tl_bal_gt_0",
    "num_sats", "num_tl_120dpd_2m", "num_tl_30dpd", "num_tl_90g_dpd_24m",
    "num_tl_op_past_12m", "pct_tl_nvr_dlq", "percent_bc_gt_75", "pub_rec_bank
    "tax_liens", "tot_hi_cred_lim", "total_bal_ex_mort", "total_bc_limit",
    "total_il_high_credit_limit", "hardship_flag", "debt_settlement_flag"
]

target = ["loan_status"]
```

In [4]:
```python
# Load the data
file_path = Path('LoanStats_2019Q1.csv.zip')
df = pd.read_csv(file_path, skiprows=1)[:-2]
df = df.loc[:, columns].copy()

# Drop the null columns where all values are null
df = df.dropna(axis='columns', how='all')

# Drop the null rows
df = df.dropna()

# Remove the `Issued` loan status
issued_mask = df['loan_status'] != 'Issued'
df = df.loc[issued_mask]

# convert interest rate to numerical
df['int_rate'] = df['int_rate'].str.replace('%', '')
df['int_rate'] = df['int_rate'].astype('float') / 100


# Convert the target column values to low_risk and high_risk based on their v
x = {'Current': 'low_risk'}
df = df.replace(x)

x = dict.fromkeys(['Late (31-120 days)', 'Late (16-30 days)', 'Default', 'In
df = df.replace(x)

df.reset_index(inplace=True, drop=True)

df.head()
```

Out[4]:

	loan_amnt	int_rate	installment	home_ownership	annual_inc	verification_status	issue_d
0	10500.0	0.1719	375.35	RENT	66000.0	Source Verified	Mar-2019
1	25000.0	0.2000	929.09	MORTGAGE	105000.0	Verified	Mar-2019
2	20000.0	0.2000	529.88	MORTGAGE	56000.0	Verified	Mar-2019
3	10000.0	0.1640	353.55	RENT	92000.0	Verified	Mar-2019
4	22000.0	0.1474	520.39	MORTGAGE	52000.0	Not Verified	Mar-2019

5 rows × 86 columns

Split the Data into Training and Testing

In [5]:
```
# Create our features
X = df_binary_encoded.drop(columns="loan_status", axis=1)


# Create our target
y = df["loan_status"]
y
```

In [6]:
```
X.describe()
```

Out[6]:

	loan_amnt	int_rate	installment	annual_inc	dti	delinq_2yrs	i
count	68817.000000	68817.000000	68817.000000	6.881700e+04	68817.000000	68817.000000	
mean	16677.594562	0.127718	480.652863	8.821371e+04	21.778153	0.217766	
std	10277.348590	0.048130	288.062432	1.155800e+05	20.199244	0.718367	
min	1000.000000	0.060000	30.890000	4.000000e+01	0.000000	0.000000	
25%	9000.000000	0.088100	265.730000	5.000000e+04	13.890000	0.000000	
50%	15000.000000	0.118000	404.560000	7.300000e+04	19.760000	0.000000	
75%	24000.000000	0.155700	648.100000	1.040000e+05	26.660000	0.000000	
max	40000.000000	0.308400	1676.230000	8.797500e+06	999.000000	18.000000	

8 rows × 95 columns

In [7]:
```
# Check the balance of our target values
y.value_counts()
```

Out[7]:
```
low_risk     68470
high_risk      347
Name: loan_status, dtype: int64
```

In [8]:
```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1, str
X_train.shape
```

Oversampling

In this section, you will compare two oversampling algorithms to determine which algorithm results in the best performance. You will oversample the data using the naive random oversampling algorithm and the SMOTE algorithm. For each algorithm, be sure to complete the folliowing steps:

1. View the count of the target classes using `Counter` from the collections library.
2. Use the resampled data to train a logistic regression model.
3. Calculate the balanced accuracy score from sklearn.metrics.
4. Print the confusion matrix from sklearn.metrics.
5. Generate a classication report using the `imbalanced_classification_report` from imbalanced-learn.

Note: Use a random state of 1 for each sampling algorithm to ensure consistency between tests

Naive Random Oversampling

In [9]:
```
# Resample the training data with the RandomOversampler
from imblearn.over_sampling import RandomOverSampler

ros = RandomOverSampler(random_state=1)
X_resampled, y_resampled = ros.fit_resample(X_train, y_train)

Counter(y_resampled)
```

Out[9]:
```
Counter({'loan_status': 1})
```

In [10]:
```
# Train the Logistic Regression model using the resampled data
from sklearn.linear_model import LogisticRegression

model = LogisticRegression(random_state=1)
model.fit(X_resampled, y_resampled)
```

Out[10]:
```
LogisticRegression(random_state=1)
```

In [ ]:
```
y_pred = model.predict(X_test)

results = pd.DataFrame({"Prediction": y_pred, "Actual": y_test}).reset_index(
results.head(20)
```

In [11]:
```
# Calculated the balanced accuracy score
from sklearn.metrics import accuracy_score

acc_score = (accuracy_score(y_test, y_pred))
print(accuracy_score(y_test, y_pred))
```

Out[11]:
```
0.646602844334948
```

In [12]:
```
# Display the confusion matrix
from sklearn.metrics import confusion_matrix, classification_report

matrix = confusion_matrix(y_test, y_pred)
cm_df = pd.DataFrame(
    matrix, index=["Actual High-Risk", "Actual Low-Risk"], columns=["Predicte
cm_df
```

Out[12]:
```
array([[  75,   26],
       [7686, 9418]])
```

In [13]:
```
# Print the imbalanced classification report
from imblearn.metrics import classification_report_imbalanced
print(classification_report_imbalanced(y_test, y_pred))
```

```
                   pre       rec       spe        f1       geo       iba
sup

  high_risk       0.01      0.74      0.55      0.02      0.64      0.42
101
   low_risk       1.00      0.55      0.74      0.71      0.64      0.40
17104

avg / total       0.99      0.55      0.74      0.71      0.64      0.40
17205
```

SMOTE Oversampling

In [14]:
```
# Resample the training data with SMOTE
from imblearn.over_sampling import SMOTE

X_resample2, y_resample2 = SMOTE(random_state=1, sampling_strategy='auto').fi
```

Out[14]:
```
Counter({'loan_status': 1})
```

In [15]:
```
# Train the Logistic Regression model using the resampled data
model = LogisticRegression(random_state=1)

model.fit(X_resample2, y_resample2)
y_pred_sm = model.predict(X_test)
```

Out[15]:
```
LogisticRegression(random_state=1)
```

In [16]:
```
# Calculated the balanced accuracy score
from sklearn.metrics import balanced_accuracy_score

acc_score2 = balanced_accuracy_score(y_test, y_pred_sm)
acc_score2
```

Out[16]:
```
0.662394124702461
```

In [17]:
```
# Display the confusion matrix
matrix_sm = confusion_matrix(y_test, y_pred_sm)

cm2_df = pd.DataFrame(
    matrix_sm, index=["Actual High-Risk", "Actual Low-Risk"], columns=["Predi
cm2_df
```

Out[17]:
```
array([[   64,    37],
       [ 5283, 11821]])
```

In [18]:
```
# Print the imbalanced classification report
print(classification_report_imbalanced(y_test, y_pred_sm))
```

```
                   pre       rec       spe        f1       geo       iba       
sup

  high_risk       0.01      0.63      0.69      0.02      0.66      0.44       
101
   low_risk       1.00      0.69      0.63      0.82      0.66      0.44     
17104

avg / total       0.99      0.69      0.63      0.81      0.66      0.44     
17205
```

Undersampling

In this section, you will test an undersampling algorithms to determine which algorithm results in the best performance compared to the oversampling algorithms above. You will undersample the data using the Cluster Centroids algorithm and complete the folliowing steps:

1. View the count of the target classes using `Counter` from the collections library.
2. Use the resampled data to train a logistic regression model.
3. Calculate the balanced accuracy score from sklearn.metrics.
4. Print the confusion matrix from sklearn.metrics.
5. Generate a classication report using the `imbalanced_classification_report` from imbalanced-learn.

Note: Use a random state of 1 for each sampling algorithm to ensure consistency between tests

In [19]:
```
# Resample the data using the ClusterCentroids resampler
# Warning: This is a large dataset, and this step may take some time to compl
from imblearn.under_sampling import ClusterCentroids

cc = ClusterCentroids(random_state=1)
X_resample3, y_resample3 = cc.fit_resample(X_train, y_train)
Counter(y_resample3)
```

Out[19]: `Counter({'loan_status': 1})`

In [20]:
```
# Train the Logistic Regression model using the resampled data
from sklearn.linear_model import LogisticRegression

model = LogisticRegression(random_state=1)
model.fit(X_resample3, y_resample3)
y_pred_cc = model.predict(X_test)
```

Out[20]: `LogisticRegression(random_state=1)`

In [21]:
```python
# Calculated the balanced accuracy score
from sklearn.metrics import balanced_accuracy_score

acc_score3 = balanced_accuracy_score(y_test, y_pred_cc)
acc_score3
```

Out[21]:
```
0.5442166848817717
```

In [22]:
```python
# Display the confusion matrix
from sklearn.metrics import confusion_matrix
matrix_cc = confusion_matrix(y_test, y_pred_cc)

cm3_df = pd.DataFrame(
    matrix_cc, index=["Actual High-Risk", "Actual Low-Risk"], columns=["Predi
cm3_df
```

Out[22]:
```
array([[   68,    33],
       [10003,  7101]])
```

In [23]:
```python
# Print the imbalanced classification report
from imblearn.metrics import classification_report_imbalanced
print(classification_report_imbalanced(y_test, y_pred_cc))
```

```
                   pre       rec       spe        f1       geo       iba
sup

  high_risk       0.01      0.67      0.42      0.01      0.53      0.29
101
   low_risk       1.00      0.42      0.67      0.59      0.53      0.27
17104

avg / total       0.99      0.42      0.67      0.58      0.53      0.27
17205
```

Combination (Over and Under) Sampling

In this section, you will test a combination over- and under-sampling algorithm to determine if the algorithm results in the best performance compared to the other sampling algorithms above. You will resample the data using the SMOTEENN algorithm and complete the folliowing steps:

1. View the count of the target classes using `Counter` from the collections library.
2. Use the resampled data to train a logistic regression model.
3. Calculate the balanced accuracy score from sklearn.metrics.
4. Print the confusion matrix from sklearn.metrics.
5. Generate a classication report using the `imbalanced_classification_report` from imbalanced-learn.

Note: Use a random state of 1 for each sampling algorithm to ensure consistency between tests

In [24]:
```python
# Resample the training data with SMOTEENN
# Warning: This is a large dataset, and this step may take some time to compl
from imblearn.combine import SMOTEENN
smote_enn = SMOTEENN(random_state=1)
X_resample4, y_resample4 = smote_enn.fit_resample(X, y)
```

Out[24]:
```
Counter({'loan_status': 1})
```

In [25]:
```python
# Train the Logistic Regression model using the resampled data
from sklearn.linear_model import LogisticRegression
model = LogisticRegression(random_state=1)

model.fit(X_resample4, y_resample4)
from sklearn.metrics import confusion_matrix
y_pred_st = model.predict(X_test)
```

Out[25]:
```
LogisticRegression(random_state=1)
```

In [26]:
```python
# Calculated the balanced accuracy score
from sklearn.metrics import balanced_accuracy_score
acc_score4 = balanced_accuracy_score(y_test, y_pred_st)
acc_score4
```

Out[26]:
```
0.6400726134353378
```

In [27]:
```python
# Display the confusion matrix
from sklearn.metrics import confusion_matrix
matrix_st = confusion_matrix(y_test, y_pred_st)

cm4_df = pd.DataFrame(matrix_st, index=["Actual High-Risk", "Actual Low-Risk"
cm4_df
```

Out[27]:
```
array([[   71,    30],
       [7232, 9872]])
```

In [28]:
```python
# Print the imbalanced classification report
from imblearn.metrics import classification_report_imbalanced
print(classification_report_imbalanced(y_test, y_pred_st))
```

```
                   pre       rec       spe        f1       geo       iba
sup

  high_risk       0.01      0.70      0.58      0.02      0.64      0.41
101
   low_risk       1.00      0.58      0.70      0.73      0.64      0.40
17104

avg / total       0.99      0.58      0.70      0.73      0.64      0.40
17205
```

In [3]:
```python
# by Emmanuel Martinez
```