Overview of the Analysis

In this analysis, we use various techniques to train and evaluate models based on loan risk. We used a dataset of historical lending activity from a peer-to-peer lending services company to build a model that can identify the creditworthiness of borrowers. Using a machine learning model, we try to determine which loans are healthy (low-risk) versus which are high-risk based on the loan status provided by the lending company.

Logistic regression is a statistical method used to predict the outcome of a dependent (loan status) variable based on previous observations. It's a type of regression analysis and is a commonly used algorithm for solving binary classification problems. Here using logistic regression makes sense as we are trying to predict whether a loan is high risk or low risk.

The dependent variable (y) in our dataset was imbalanced, meaning that most of the data belongs to non-risky (0) class (n=75036), while the risky class (1) only had 2500 observations.

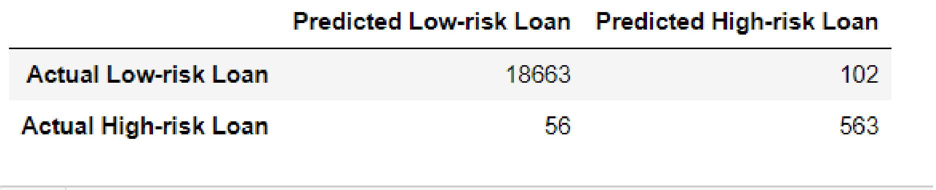
```
y.value_counts()
# output
0   75036
1    2500
Name: loan_status, dtype: int64
```

After choosing the model, we first trained the model. We passed our data through the model to find patterns and make predictions. Then we evaluated the model. This is done by testing the model on previously unseen data. Then we used the model to make predictions.

Since our data was imbalanced, we used random over sampling method to resample the data. Random oversampling involves randomly selecting examples from the minority class (here 1) and adding them to the training dataset.

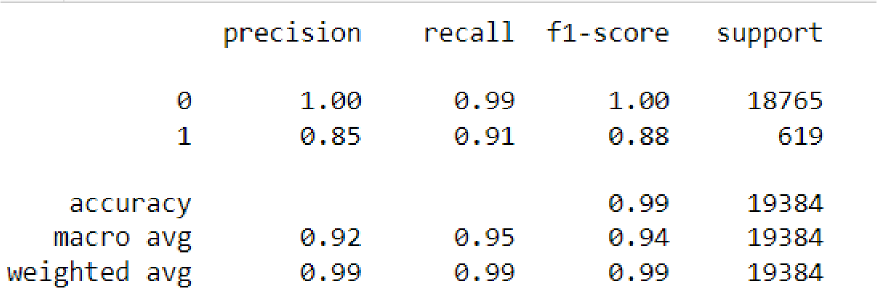
Results

* Machine Learning Model 1: Logistic Regression Model



	Predicted Low-risk Loan	Predicted High-risk Loan
Actual Low-risk Loan	18663	102
Actual High-risk Loan	56	563

From this matrix we can see that 56 actual high-risk loans were predicted as low risk as per this model. While 102 low risk loans were predicted as high risk.



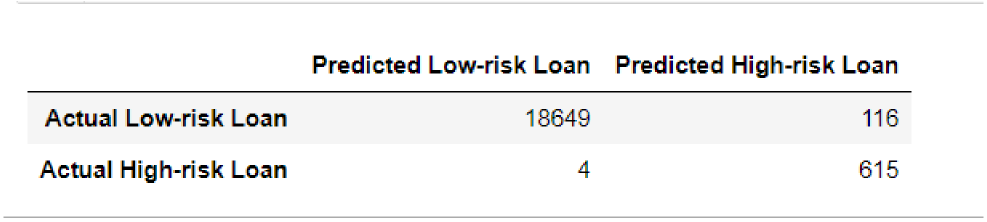
```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00     18765
           1       0.85      0.91      0.88       619

    accuracy                           0.99     19384
   macro avg       0.92      0.95      0.94     19384
weighted avg       0.99      0.99      0.99     19384
```

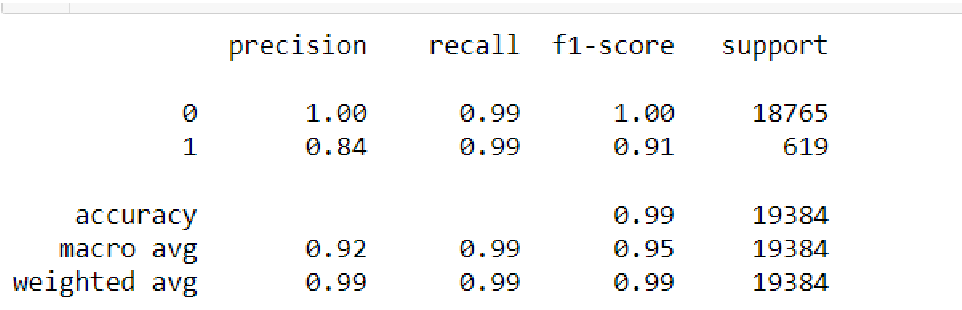
Using the dataset provided by the lending company, we created a Logistic Regression Model that generated an accuracy score of 95%. Although the model generated a high accuracy, the model's recall value (0.91) for non-healthy loans is lower than the recall value (0.99) for healthy loans. This indicates that the model will predict loan status's as healthy better than being able to predict loan status's as non-healthy. This is due to the dataset being imbalanced, meaning that most of the data belongs to non-risky (0) class (n=75036) compared to the risky class (1) with only 2500 observations. In situation like this where we have an uneven class distribution, F1 score is usually more useful than accuracy. Here the F1 score for score risky loans (1) is 0.88. This means that the model's ability to both capture high- risk loan and be accurate with the cases it does capture is 0.88.

* Machine Learning Model 2: Logistic Regression with Random Over Sampling Model



	Predicted Low-risk Loan	Predicted High-risk Loan
Actual Low-risk Loan	18649	116
Actual High-risk Loan	4	615

From this matrix we can see that only 4 actual high-risk loans were predicted as low risk as per this model. While 116 low risk loans were predicted as high risk.



```
              precision    recall  f1-score   support

           0       1.00      0.99      1.00     18765
           1       0.84      0.99      0.91       619

    accuracy                           0.99     19384
   macro avg       0.92      0.99      0.95     19384
weighted avg       0.99      0.99      0.99     19384
```

In this model our accuracy increased to 99%. Also, the recall values for both high risk and low risk loans were 0.99. This indicates that that the model will predict both types of loans equally well. In this model the F1 score for the high-risk class was better than the 1st model. The model's ability to both capture high- risk loan and be accurate with the cases it does capture is 0.91.

Summary

The Logistic Regression model fitted with Oversampled data (model 2) performed better than the model 1 fitted with imbalanced data generating a higher accuracy score and a higher recall, indicating that the model will make fewer mistakes when classifying non-healthy loans. Also, the F1 score for the second model for predicting high risk loan was 0.91 compared to 0.88 from model 1.

The lending company would want fewer False Positives due to the possibility of the lending company losing money by classifying non-healthy loans as healthy. In our analysis, even though the False Negatives went up for Model 2 compared to Model 1, it is aways better to err on the side of caution. It is better to predict a healthy loan to be non-healthy and not lend money compared to the other scenario where the lending entity will lose money by giving out a bad loan.

Model 1 fitted with Imbalanced Data:

102 FALSE NEGATIVES --> The actual value is low-risk and the predicted value is high-risk

56 FALSE POSITIVES --> The actual value is high-risk and the predicted value is low-risk

Model 2 fitted with Balanced Data:

116 FALSE NEGATIVES --> The actual value is low-risk and the predicted value is high-risk

4 FALSE POSITIVES --> The actual value is high-risk and the predicted value is low-risk

According to the confusion matrices, the number of False Positives decreases indicating the model will classify non-healthy loans as risky better. Based off this analysis, I would recommend using Model 2 (Logistic Regression with Random Over Sampling Model).