**Lending Club Loan Data Analysis**

DESCRIPTION

Create a model that predicts whether or not a loan will be default using the historical data.

Problem Statement:

For companies like Lending Club correctly predicting whether or not a loan will be a default is very important. In this project, using the historical data from 2007 to 2015, you have to build a deep learning model to predict the chance of default for future loans. As you will see later this dataset is highly imbalanced and includes a lot of features that makes this problem more challenging.

Domain: Finance

Analysis to be done: Perform data preprocessing and build a deep learning prediction model.

Steps to perform:

Perform exploratory data analysis and feature engineering and then apply feature engineering. Follow up with a deep learning model to predict whether or not the loan will be default using the historical data.

Tasks:

1. Feature Transformation

Transform categorical values into numerical values (discrete)

1. Exploratory data analysis of different factors of the dataset.
2. Additional Feature Engineering

You will check the correlation between features and will drop those features which have a strong correlation

This will help reduce the number of features and will leave you with the most relevant features

1. Modeling

After applying EDA and feature engineering, you are now ready to build the predictive models

In this part, you will create a deep learning model using Keras with Tensorflow backend

STEPS Performed:

1. Import the namespaces
2. Load the Loan Dataset.
3. Exploratory data analysis – check the shape and size of data, check the no.of columns, check for null data etc.
4. Notice that the y variable “not.fully.paid” is high imbalanced.

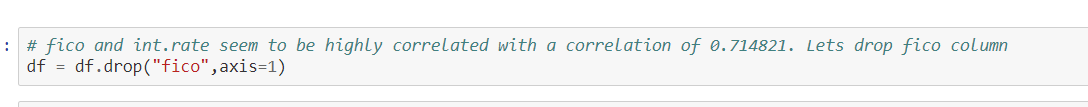
Chart, bar chart

Description automatically generated

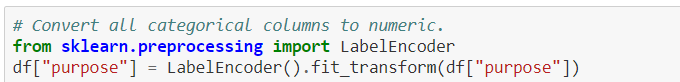
Graphical user interface, text, application

Description automatically generated

1. Find the correlation between variables. Drop the variables that are highly correlated.

6. Perform EDA on the data.

7. Convert categorical to numeric.



8. Define X and y variables.

9. Scale the X variables.

10. Split the data into train and test.

11. Architect the model.

Text

Description automatically generated

12. Compile and train the model.

13. Plot the train and validation accuracy and errors.

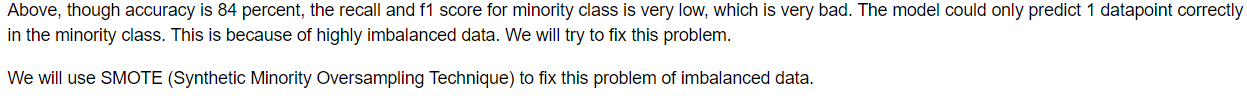
Graphical user interface, application

Description automatically generated

14. Check the confusion matrix. Notice that though accuracy is good, f1 score and recall for minority class are poor.

Table

Description automatically generated



15. Use SMOTE to over sample the minority class.

Text

Description automatically generated

Graphical user interface, text, application, chat or text message

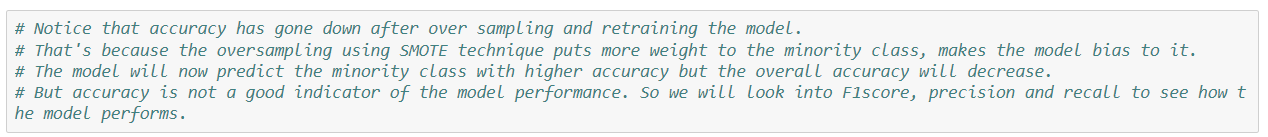
Description automatically generated

16. Train the model again with the over sampled data.

17. Plot the train and validation accuracy.

Graphical user interface, chart

Description automatically generated



18. Plot the threshold for True Positives and False Positives. Draw the ROC Curve and find the optimal threshold.

Graphical user interface, chart

Description automatically generated

19. Check the confusion matrix. Notice that precision, recall and f1 scores for minority class has improved after addressing the imbalance in dataset issue.

Chart, treemap chart

Description automatically generated

Text

Description automatically generated

Conclusion:

With unbalanced data, the f1 score and recall for minority class were very low around 0.04.

We tried to fix the unbalanced dataset issue by using SMOTE. SMOTE uses KNN algorithm to generate synthetic records. This is called over sampling. We used this technique to over sample the minority class records. After oversampling and re-training, the model, the model performs better in classifying the minority class. Notice that accuracy has gone down after over sampling and retraining the model. That's because the oversampling puts more weight to the minority class, makes the model bias to it. The model will now predict the minority class with higher accuracy, but the overall accuracy will decrease.