GyaanData Hiring Assignment

By: Ranjit Nair

Overview

We are given a dataset, that prima facie looks like a peer to peer lending dataset, which spans multiple countries and languages in a csv format. The task at hand is to do exploratory data analysis as much as possible, get some valuable insights from the data, and then build a Machine Learning model that tells us with high precision and high specificity whether a particular loan application needs to be funded or not.

Assumptions

1 We assume that the data given to us is a peer to peer lending data mainly because of the description of the variables.

2 In solving the given Machine Learning Problem Statement, we assume that the given data accurately describes the funding status of a loan application and that there is negligible default among approved loans

3 We also assume in the end that correctly classifying a loan application has some benefits and incorrectly classifying leads to some loss but that information is not known to us at the moment.

Project objective

To develop a Machine Learning algorithm that classifies the funding status of a particular loan application with high Precision (True Positive Rate)and also high Specificity (True Negative Rate).

Some basic Insights about the Data before we build the Model.

The nature of DataSet

We have close to 55000 data points about various loan applications and their funding status. We're also given 14 columns variables like:

1) Activity
2) Gender of the Borrower
3) Country
4) Country Code
5) Currency of the Policy
6) Distribution Model
7) Lender Count
8) Loan Amount
9) Original Language
10) Repayment Interval
11) Sector
12) Term of Loan in Months
13) Multi Poverty Index
14) Funding Status of the Loan

Top 10 type of activities for which loan was requested.

1.	General Store:	4771
2.	Farming:	4655
3.	Retail:	3009
4.	Clothing Sales:	2559
5.	Personal Housing Expenses:	2550
6.	Food Production/Sales:	2247
7.	Agriculture:	1831
8.	Grocery Stores:	1714
9.	Home Appliances:	1582
10.	Pigs:	1449

Distribution of Loan by Gender or Group

1.	Female:	33698
2.	Male:	13178
3.	Group:	8186

Top 10 Countries in which loan was requested.

1.	Philippines:	9654
2.	Cambodia:	5166
3.	Kenya:	5130
4.	Nicaragua:	3755
5.	Peru:	3655
6.	El Salvador:	3321
7.	Tajikistan:	2159
8.	Colombia:	1941
9.	Uganda:	1837
10.	Nigeria:	1814

Distribution of Loan by Currency Policy

1.	Shared:	42908
2.	Not Shared:	12154

We see that the majority of the Loan’s Currency Policy is shared in nature.

Types of Loan Distribution Models

1. Field Partner: 55062

We can see that there is only one model of distribution of Loan, that is, Field Partner. Which gives us idea that all the data-points in the dataset are in fact of a peer to peer lending case study.

Top ten most popular number of lenders

1.	8 Lenders:	2525
2.	7 Lenders:	2437
3.	10 Lenders:	2389
4.	9 Lenders:	2385
5.	5 Lenders:	2365
6.	6 Lenders:	2278
7.	11 Lenders:	1972
8.	12 Lenders:	1948
9.	4 Lenders:	1928
10.	13 Lenders:	1813

Distribution of Languages in which Loan Form was filled

1.	English:	34525
2.	Spanish:	15911
3.	French:	3046
4.	Russian:	1280
5.	Portuguese:	118
6.	Vietnamese:	99
7.	Indonesian:	59
8.	Arabic:	24

Distribution of Sectors in which Loans were Disbursed

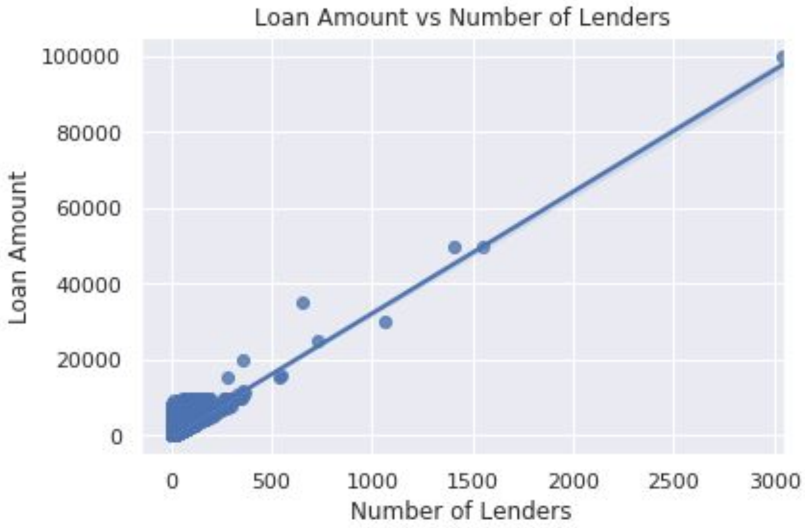
1.	Food:	12161
2.	Retail:	11944
3.	Agriculture:	11439
4.	Services:	3821
5.	Clothing:	3443
6.	Personal Use:	2941
7.	Housing:	2691
8.	Education:	1883
9.	Transportation:	1555
10.	Arts:	1038
11.	Construction:	806
12.	Health:	618
13.	Manufacturing:	566
14.	Wholesale:	78
15.	Entertainment:	78

Plot between Loan Amount and Number of Lenders

We can see that the loan amount and the number of lenders are highly correlated.

The Correlation Coefficient = **86.8 %**

The solid line is the regression line fitted through the data points.



The Average Rate of Funding

1. The average rate of funding a loan is equal to **87.73%.**
2. We can also interpret it as the probability of a loan being funded given no other information
3. As we will see in the upcoming slides, the probability of funding will vary quite significantly across the demographics
4. It’s also apparent that the dataset is quite imbalanced when it comes to funding. We need to take care of this as most Machine Learning Algorithms don’t perform well on Imbalanced datasets

Funding percentage for major Languages.

1.	French: 94.2%	(above average)
2.	English: 90.8%	(above average)
3.	Spanish: 80.9%	(below average)
4.	Russian: 73.2%	(below average)

We can see that French and English have a better average funding rate whereas Spanish and Russian language have a considerably below the average funding rate of 87.73%

Funding percentage for various Genders or Groups

1. Female: 92.0%	(above average)
2. Group: 91.0%	(above average)
3. Male: 74.7%	(below average)

We can see that Females and Group have a better average funding rate whereas Males have a considerably below the average funding rate of 87.73%

Funding percentage for Major Sectors

1. Arts Sector: 99.7%	(above average)
2. Education Sector: 93.2%	(above average)
3. Food Sector: 91.2%	(above average)
4. Personal Use: 89.4%	(above average)
5. Services Sector: 88.2%	(above average)
6. Agriculture Sector: 87.5%	(below average)
7. Retail Sector: 85.5%	(below average)
8. Clothing Sector: 84.0%	(below average)
9. Transportation Sector: 82.8%	(below average)
10. Housing Sector: 73.2%	(below average)

Funding Percentage for Major Activities

1. Home Appliances: 97.7%	(above average)
2. Food Production/Sales: 91.8%	(above average)
3. Farming: 88.4%	(above average)
4. General Store: 87.3%	(below average)
5. Agriculture: 85.1%	(below average)
6. Retail: 85.0%	(below average)
7. Clothing: 81.6%	(below average)

Funding percentages for various Repayment Intervals

1. Irregular Term: 95.6% (above average)
2. Monthly Term: 84.3% (below average)
3. Bullet Term: 80.0% (below average)

Funding Percentage for Major Countries

1. Philippines: 97.0%	(above average)
2. Cambodia: 95.3%	(above average)
3. Peru: 95.0%	(above average)
4. Kenya: 86.2%	(below average)
5. Nicaragua: 84.8%	(below average)
6. Tajikistan: 76.0%	(below average)
7. El Salvador: 68.3%	(below average)

Performance of a Dummy Classifier

1. We start off with a Dummy Classifier that just categorizes every loan application as funded.
2. Even though the Dummy Classifier is not using any Machine Learning technique, it's accuracy is about **87.72%,** which is quite high for an algorithm which basically does nothing.
3. This happens because we have an imbalanced dataset and for an imbalanced dataset accuracy is not a good metric to analyze its performance as can be seen in the previous point.
4. Thus we disregard accuracy and instead use two performance metrics, namely:
    a. Precision: Otherwise known as The True Positive Rate. **[tp / (tp + fp)]**
    b. Specificity: Otherwise known as The True Negative Rate. **[tn / (tn + fn)]**
5. Next we'll see how an out of the box K-Nearest Neighbors Classifier performs.

Performance of K-Nearest Neighbors

1. Here we use K-Nearest Neighbors which is one of the simplest Machine Learning Algorithm.
2. The KNN Classifier has a Precision of: **95.34%**
3. The KNN Classifier has a Specificity of: **69.75%**
4. We see that the KNN Classifier performs decently when it comes to Precision but has a sub-optimal performance when it comes to Specificity.

Performance of Random Forest

1. Here we use Random Forests which is one of the most used ensemble Machine Learning Algorithm.
2. The Random Forest Classifier has a Precision of: **94.86%**
3. The Random Forest Classifier has a Specificity of: **82.85%**
4. We see that the Random Forest Classifier performs overall performs better than K-Nearest Neighbors Classifier because:
    a. It underperforms KNN Classifier on Precision by merely **0.48%**
    b. But it outperforms KNN Classifier on Specificity by a huge **13.10%**
5. We've yet to address the imbalance in the dataset which generally degrades the performance of the Machine Learning Algorithms.
6. In order to address the imbalance, we oversample from the minority class using a technique called SMOTE which stands for Synthetic Minority Over-Sampling Technique and use it in combination with Random Forest Classifier.

Performance of Random Forests + SMOTE

1. Here we address the imbalance in the DataSet and use Random Forest Classifier in conjunction with SMOTE (Synthetic Minority Over-sampling Technique)
2. The Random Forest Classifier in conjunction with SMOTE has a Precision of: **98.71%**
3. The Random Forest Classifier in conjunction with SMOTE has a Specificity of: **95.81%**
4. We see that the Random Forest Classifier in conjunction with SMOTE performs much better than pure Random Forest Classifier as:
    a. It outperforms pure Random Forest Classifier on Precision by a significant **3.85%**
    b. It outperforms pure Random Forest Classifier on Specificity by a huge **12.96%**
5. In the next slide we'll see how we can further improve the performance of our model and the way forward.

The Way Forward

1. The previous model could be further enhanced by tuning HyperParameters using Grid Search or Randomized Grid Search mechanism but given the limited time and computational resources, it wasn’t attempted in this case study. It is expected that by tuning hyperparameters, one can further improve the performance metric by a few percentage points hopefully.
2. Till now we’ve only considered the technical side of the problem and completely ignored the business side of things. If we are given a cost matrix of benefits and losses corresponding to correct classification and misclassification respectively, we can dot product it with the confusion matrix produced by the model and get a better metric to optimize for than the theoretical Precision and Specificity. For this, we would need the business cost and profits due to correct or incorrect classifications, which wasn’t provided in this case. Incorporating that would lead to building a more interpretable Machine Learning Model.

Thank you.