LOAN DEFAULT PREDICTION IN MICRO FINANCE BY DEEP JOSHI

OVERVIEW

Kiva is an online loan lending platform that allows individuals to make small loans to borrowers across the world.

Our analysis aims at studying loan defaults across countries having a decent mixture of defaulted and paid loans. Thus, I have chosen a sample of observations from countries like Ecuador, Kenya and Dominican Republic as they have decent number of observations of defaulted loans with respect to paid loans.

Finally, I modelled the data after processing it and constructed 3 decision tree classifiers:

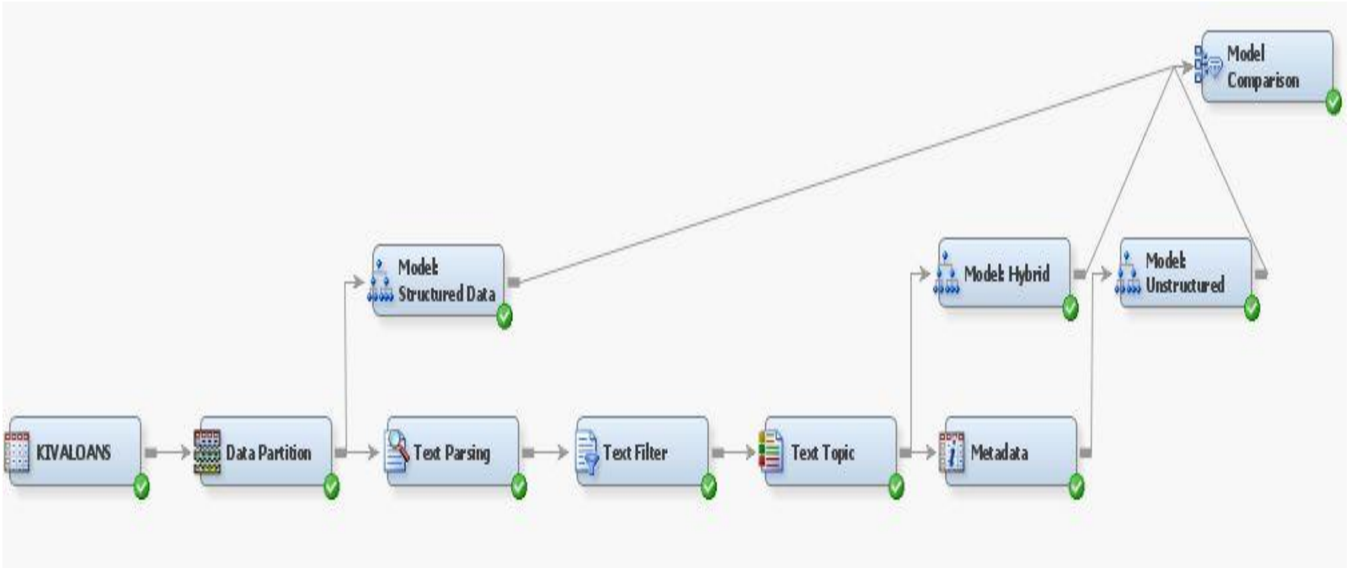
- Decision Tree Classifier with Structured Data only:
  - Here, I have directly trained the dataset in the tabular format using the decision tree classifier

- Decision Tree Classifier with Unstructured Data only:
  - In this method, I have created a text mining flow to process the textual data contained in the relational dataset.
  - Several NLP tasks like parsing, stop words removal, tokenization, stemming, parts of speech tagging and spell checking, were performed on the textual data before finally generating text topics.

- Hybrid Model: (Structured + Unstructured Data):
  - In this method, I have processed the data up till text topic stage but instead of training the model just on text data, I trained all the features together.

Finally, I compared the performance of all the tree models to select the best one based on various metrics.

FLOW DIAGRAM



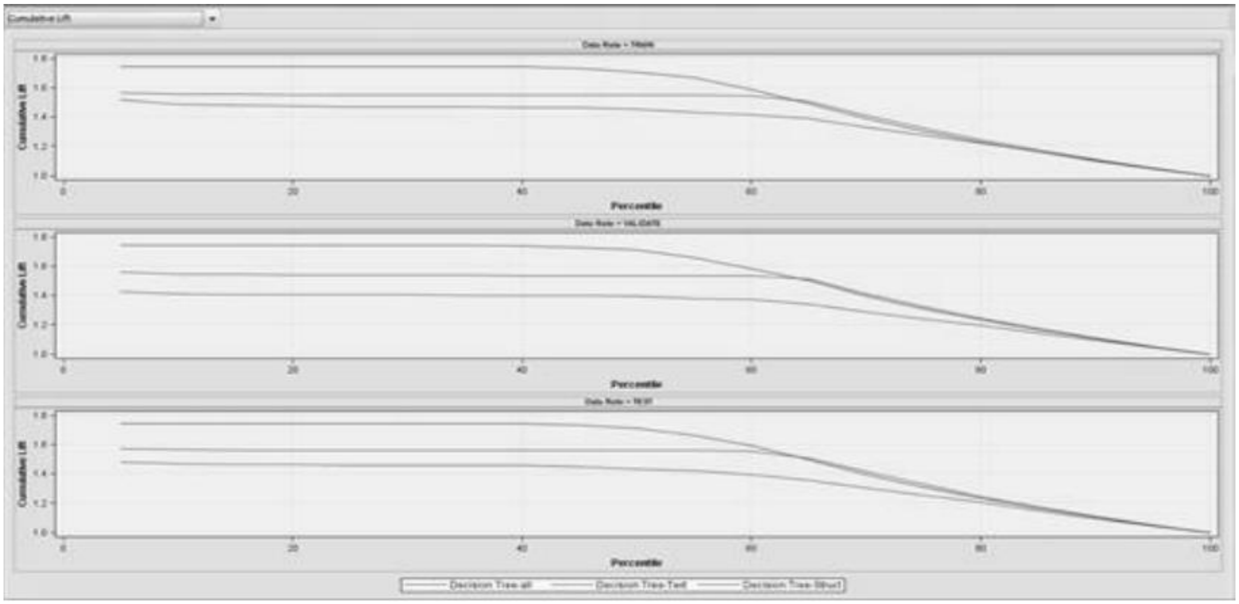
DATASET:

The dataset contains loan information that is translated from other languages and contains verbiage unique to country and religion. Thus, in such cases a text mining flow plays a pivotal role in mining accurate information from the data.

Our dataset consists of various features like

1. Non-Payment: Describes the entity that takes on the risk (lender/bank/partner institution).
    - Based on the data and as per my analysis lender takes more risks.
2. Country – Kenya, Ecuador or Dominican Republic
3. Loan Amount
    - It was found that borrowers with larger loan amounts tend to pay the amount in most of the cases i.e. probability of repayment is high when the loan amount is high
4. En – States the purpose of why the loan was borrowed
5. Gender – Male or female
6. Sector – transportation, business, education, etc.
7. Status (Target) – Default or not

COMPARISION



Comparing the models over various performance metrics:

1. Text only performed the worst
2. Relational only performed better
3. Hybrid model outperformed both

INSIGHTS

When I studied the output of these three models all together, third model won the race i.e. when trained on structured and unstructured data all together. The runner up was the model trained on completely relational data and the model trained on text data only, performed the worst.

The reason why the text only data performed the worst is due to the fact the data is quite dirty. Firstly, it has been translated from several other languages into English. Secondly, it has so much noise that require manual intervention and large number of iterations to clear it. In this model, there were no significant differentiators contributing to the defaults and no defaults.

Introducing relational data into the picture highlighted me with several good features that are not only good predictors, but also allowed to understand the audience who make into the defaulter list and those who not.

By analyzing the behavior of introducing textual data along with tabular data into our model, I found that people who borrowed the loans for purchasing something or growing, had a higher propensity of paying the loan back. Textual data defining the purpose of loan borrow definitely added value to the model. Also, I figured out one more interesting fact – textual data weighed more than the country in the prediction task.

When the cumulative effect of these models was studied, structured only and text data only performed almost similar while the earlier one just edging over the later one. However, much information was generated from the third model by the introduction of processed textual content with the relational data.