Alternative Lending Platform using AI & Big data enabled Social Credit

Aviruk Chakroborty
BE Mechanical Engineering

Abstract: Today, many people in both first world and emerging economy do not have access to capital due to insufficient credit history. This platform which uses social media as one of its parameters is a way towards providing credit to a larger population group. Since, the subprime group is not availing the general credit rating bureau scores, a more comprehensive data capture protocol from alternative sources must be deployed for creating the credit rating models. These alternative credit rating modules span across a wide user base right from subprime users in the first world to the unbanked population in emerging economies. The lending process is divided into two parts, Direct user flow to the portal where post submission of their details which includes sharing of certain basic information is permissions like social media field, bank account details and phone data use permission. A robust credit rating module is built.

Introduction

This platform also employs a proactive user capture protocol which is explained below along with the entire process map:

1. A custom data capture spider including separate email capture spiders are deployed over a target geography. This include data scraping from various deep web applications also which includes, duck duck go. These data points along with data mining applications and scrapers which scrap all the company data available in that country, region and various other openly available public sources of data are scrapped and mined. We also mine exhaustively on the internet, all the information publicly available about each applicant - both retail and corporate.

2. User profiling using artificial intelligence and machine learning in form of demographic bucketing and natural language processing. The NLP enabled AI is able to contextualise the data mined and segregate them in various user profile buckets. For example, we bucket users’ into very refined socio economic demographics like
   a. Father with one car
   b. Student in pen ultimate year
   c. Corporate entities
   d. Factories
   e. Single mothers
   f. Housewives
   g. Socially mobile single high earners
   h. Many more such targeted profiles.

Here we send them targeted campaigns like to a pen ultimate students - we send education loan profiles, To a housewife, we send a medical insurance for family, for fathers’ with children, we send them auto loans.

3. Automated communication protocol: Post socio economic bucketing using NLP, we cross-reference the user data with email spiders and various other data mined resources which include phone number and automated messages and mails and calls are dispatched with a single objective of bringing the user into the online lending portal.

4. Once the user registers, we run a series of alternative data scraping protocols using which we are able to create a credit model. These include but are not limited to the following:
   a. Duck Duck go based deep web search
   b. Electronic KYC
   c. Neural net based document verification
   d. Fraud analytics
   e. social media based credit rating
   f. cross-referencing with national and international databases
   g. biometric data cross validation
   h. Cell phone data
   i. cross checking with past educational institutions automatically via a script
   j. Industry Ministry and registrar of companies for cross validating employers and places of employment
k. Cross validation of each data field including probability distribution based cross check of the default probability using fraud analytics. For example, during e-KYC process, a couple of data points like driver's license, Aadhaar/Social security card are provided by loan applicants to the platform. Using these data points, we cross-validate as many fields as possible which are filled by the user. In emerging economies, data validations is one of the major pain points as data is not available for creation of validation models.

6. Application of AI models like Face recognition, face detection, sentiment analysis etc. on the gathered data. This also includes running a social media based credit rating analysis where the output i.e. credit score becomes one of the decision factors in step 7 for dispersal of credit. Through our APIs, we also ask the loan seeker to come and provide us with their baking history. Our API integrates with baking institutions and the bank statement is extracted and scored using or machine learning algorithm. We also run deep web search of the user and scrap all the text and pictures associated with that person. Then we cross validate the image folder with the users' photo and if the photo of the user is not present anywhere on the internet inspite of the user being the right demographic bucket for example, between age 25-35 with a masters degree and earning above 100000$ this becomes a potential red flag which gets transferred to our step 7 which is our lending algorithm. We also run a negative word filter and sentiment analysis using Natural Language Processing Artificial Intelligence framework on all the text scrapped about the person from all public sources to generate a likability score. The word filter also helps us identify potential new stories about the firm or the individual. This helps us immensely in updating our credit lending module.

7. Creation of alternative credit models using an ensemble classifier and using various diverse models to create and integrated hybrid lending model. In this model, we use neural networks, logistic regression, autoencoders and various other sub classifiers in the space of machine learning like SVM, Apriori, KNN, Kmeans, Hidden markov models and a collaborative and item based recommendation engine.

   a. Stage 1 is always logistic regression for credit models till we reach a higher data rich volume.

   b. Step 2 is running an autoencoder based feature extractor to objectively reclassify the most important weights.

   c. Next we use the autoencoder and heuristic weights as the initialisation weights of a back propagation based neural network.

   d. In various specific loans, like auto loans, the car IOT data stream is also factored in decision making process.

8. When there are red flags and low volume of data and the confidence score of the algorithmic rating model for lending is not high enough to warrant an automatic dispersal then the loan provider can conduct a video conference with the potential loan seeker. In this video conference, we also provide the Fund providers with various artificial intelligence tools which will help them in making decisions. These are facial recognition so that the person who has applied is the person attending the video conference, emotion quantification and sentiment analysis so that the lender has a clear idea about the emotional state of the loan seeker as this is necessary in such digital platforms where the emotional part of the decision-making is sensitised. Our AI module also calculates the emotion using voice stress analysis over the audio feed of the video conference. Today, using webcams and cellophane cameras, it is possible to measure blood pressure and pulse of the person. When a stress question is asked like ‘Will you commit fraud or not?’ pre and post that question, the delta in blood pressure, pulse, audio pitch, facial expressions are calculated and if the delta is too high a red flag is sent to the fund provider as a notification which can be used as a legitimate parameter for denying the loan. These technologies are there in nascent stage today and will become more efficient as the data depth of the models increase.

9. Using the decision framework of step 7, the master neural network which is a recurrent and LSTM based neural net, keeps updating its weights when it comes to lending models.

10. Smart contracts between the loan seekers and loan providers which is built on our own Blockchain which is derived from an Alt Coin
framework which is our proprietary cryptocurrency framework. This framework ensures immediate dispersal of money.

11. The final step of the process is monitoring module in which we integrate our system via an API into ERPs of corporate loan seekers and cell phone based credit monitoring for retail loan seekers.

Literature Review

Yann Le Cunn and others in their paper on Deep Learning explain breakthroughs that have been achieved due to CNNs. On the same basis, we have further developed several innovative applications that can change the face of finance. However, when we consider Lending Platform, it becomes very important to decide the potential of the credit seeker which is where Yanho Wei and others Credit Scoring with Social Network was used as a point of reference.

This study involves a major role of Artificial Intelligence and its various modules which were first discussed by Rosalind Picard in her study Affective Computing. The suggested directions in her paper on computing recognition of human emotion has been the key to the developed modules like Face detection and recognition which the paper discusses in detail.

Methodology

The social media based credit rating is built on two separate neural networks. Neural Net 1 is used as a classifier where based on the users' social media parameters we classify them into high, low risk buckets and socio economic buckets. The second neural net is used to learn the correlation between social media features and credit repayment propensity. This model keeps updating itself with every repayment made or not made by users to whom loan has been dispersed.

In today’s world, almost all the urban population use social media some way or the other. The presence is almost obvious. They get friends with many and react to certain posts. They write about themselves and people also write for them. They provide details of their hobbies, favourite books, favourite movies etc. All such details combined can be used to know a person’s personality traits. The presence can also provide support to their authenticity.

When a person could be found on the social media than that is a positive point of the person being a real one and not non-existent. Conversely, if a person’s name is not present on the web then there are chances that the person is fake. Moreover, what the person likes or dislikes can provide its inclination towards positive or negative things using the sentiment analysis ran on the text.

Out of all the different options, Facebook is highly used across the globe and while integrating with can provide details about below fields if the information is made public by the user. We can apply sentiment on all the information which contains texts. We can cross-reference the work experience with what the user has provided while entering the info. We can also get the list of friends and find their credit rating because a person’s behaviour and actions are many times in sync with their friends or inspired by them.

We apply logistic regression on all the details found from social media presence along with other information to calculate if the person will default or not. If the presence is not found, then the answer is neutral. Moreover, if the presence is found then each parameter will get around 0.02-0.05% weightage. That is a very small contribution however, an insightful one.

If the user installs an App on a smartphone then she can be asked to provide access to other apps in the phone from where the call logs, health kit information, GPS location, calendar based reminders, contacts etc. can be fetched. Using the location and health information, the person can be categorized of being real or fake and also gather a lot of other information.

The social media analysis also contains a network analysis in which using graph theory, we analyse the network value of the individual and derive insights from it. This network analysis also consists of credit worthiness check on the people who are part of the users' social network along with cross referencing the friends with national and financial fraud databases. This is a weighted graph theory model in which the negative weight of a friend who is closer to the person is more if that person is a declared fraudulent person. Using the graph theory analogy, the closeness between social media relationships and their public profiles also is an important factor as birds of the same feather flock together. Here, this weighted approach ensures that accepting a potential fraud as an acquaintance/social media friend does not tremendously negatively impact credit rating score. But, if the volume of contact is higher than the probability of the user not knowing the past of their acquaintance reduces in an exponential fashion.

These metrics are most useful in the age group of millennials, and will keep becoming more relevant as the volume of social media use increases. Though today, Facebook is the predominant player,
the hypothesis of this paper is social media content analysis based rating irrespective of the platform/brand as our primary target is socio economic clustering. This is especially true as 90% if the world's data was created in the past 3 years and the rate of data generation is not linear but exponential in nature.

Our integrated approach is much more robust than relying on just social media credit rating as behaviour is also reflected in the bank statements, mobile phone data and document veracity which we check using image classifiers.

These data points are anyways provided by current loan seekers to their potential providers including employment details and bank statement integration especially in the case of payday loans where the money is automatically deducted from the salary account.

While dispersing payday loans we also calculate the stability of the employer as the employer is the source of repayment. In case of public limited companies this analysis is relatively easier due to large volumetric availability of data but the difficulty keeps increasing with reduced size and digital presence of the firm. Hence, we have created a separate module for SME segment employer profiling as comparing apples and oranges is akin to over fitting the sample distribution in statistics. In payday loan, the analysis of account provides us a valuable tool in judging regularity of income. For example, in Dubai, UAE, institutions mandate minimum 5000 AED salary for loans but since Uber drivers who on an average 15000 AED per month but are contract workers by the nature of their employment contracts with the rideshare platform are not able to avail the standard lending platforms. This platform will help such unbanked users.

During evaluation of education loans, we also run social models and data scrapping from various public sources including portals like glassdoor to evaluate the average salary and proportion of unemployment post-graduation once a student graduated from a specific university and course. This is one of the most direct matrix of calculating default probability of education loans.

The application of these platforms is not limited only to online lending but can be ubiquitously be used in credit evaluation for e-commerce site purchases, insurance purchase and any other form of online commerce including hospital bill payment where a certain amount of credit has to be provided to any user for even small windows of time.

AI algorithms and AI modules, we can consider:

- ANN
- CNN
- LSTM
- Autoencoders
- Recommendation Engine
- Apriori
- Hmm
- Kmeans
- Logistics Regression
- all user actions overlayed on a time series
- bi parte graph to find the best fit between FS and FP

AI modules

We have used the following artificial intelligence modules in various formats at multiple locations of this lending supply chain, staring from application of machine learning based data mining to the use of computer vision and convolution neural nets to capture, store and predict human emotional states.

- Face Detection: We use face detection while sorting scrapped picture files from duck duck go and only pass the pictures which have faces to our facial recognition module, the same module is used on uploaded picture and the video conference module. This module is built on a machine learning enabled computer vision framework.

- Face Recognition: Using a biometric machine learning face detection framework, we cross verify the user's face with multiple data sources and update the data validity score accordingly.

- Emotion: We capture over 21 micro expressions on the face based on which we classify the human emotion into 7 basic emotional states and save this data in conjecture with a time series data framework. This technology has applications from automated bargaining in e-commerce sites to customer happiness index to customer interaction quality quantification, robotics and many more.

- Pulse: We use cell phone camera and video camera based machine learning tools which based on micro skin activity is able to infer the blood pressure and pulse of the human.

- Audio: Human speech is direct by product of the emotional state especially the frequency of the voice. Using an audio filter and machine learning classifier, we are able to classify each audio snippet into various proportional emotional states.
• Sentiment: Today, with the high volume of text generated about each user, the contextualization of text becomes the most important factor in text analysis. This is achieved with the use of Natural Language Processing which is a machine learning framework which categorises the text into various moods, sentiments like positive, neutral, negative.

• Speech to text: Using a deep neural net, we are able to convert speech to text and then run sentiment analysis on it to objectively quantify the quality of the interaction.

• Meta data-exif: Traffic analysis over the past sixty years has taught us that metadata reveals a lot more about the communication. For example, Exif filters on the pictures which are uploaded and social media and those Exif data scraps tells us about the brand of phone the user uses, the latitude and longitude and many other features. We use these metadata points to refine our user's demographics profile.

• Document verifier: We use a convolutional neural net to train on a set of valid KYC documents and then this trained CNN is used as a filter by processing all the uploaded documents. This helps us identify potential fraudulent and red flag documents. To increase the effectiveness of the CNN which includes identify low quality images, we need to increase the depth of the neural net.

• Open source macroeconomic data: We are always scrapping open source macroeconomic an industry indicated data which includes index of stock markets, forex, which is used to create predictive models on company earning which gets reflected in their credit score.

Conclusion and Road Ahead

Our integrated credit platform provides a unique opportunity to offer credit solution to the disenfranchised and help increase the credit funnel exponentially. These data models will become more and more relevant with every passing day as the volume of data is increasing exponentially. Hence, we believe that alternative credit models will be a better risk quantification measure than traditional credit rating methodologies which only look at the users' repayment behaviour in certain arenas. This approach provides a 360-degree view to the whole framework.

References