International Journal of Soft Computing and Engineering (IJSCE)
ISSN: 2231-2307 (Online), Volume-11 Issue-2, January 2022

Prediction of Credit Card Approval

Harsha Vardhan Peela, Tanuj Gupta, Nishit Rathod, Tushar Bose, Neha Sharma

Abstract*: Credit risk as the board in banks basically centers around deciding the probability of a customer's default or credit decay and how expensive it will end up being assuming it happens. It is important to consider major factors and predict beforehand the probability of consumers defaulting given their conditions. Which is where a machine learning model comes in handy and allows the banks and major financial institutions to predict whether the customer, they are giving the loan to, will default or not. This project builds a machine learning model with the best accuracy possible using python. First we load and view the dataset. The dataset has a combination of both mathematical and non-mathematical elements, that it contains values from various reaches, in addition to that it contains a few missing passages. We preprocess the dataset to guarantee the AI model we pick can make great expectations. After the information is looking great, some exploratory information examination is done to assemble our instincts. Finally, we will build a machine learning model that can predict if an individual's application for a credit card will be accepted. Using various tools and techniques we then try to improve the accuracy of the model. This project uses Jupyter notebook for python programming to build the machine learning model. Using Data Analysis and Machine Learning, we attempted to determine the most essential parameters for obtaining credit card acceptance in this project. The machine learning model we built gave an 86 % accuracy for predicting whether the credit card will be approved or not, considering the various factors mentioned in the application of the credit card holder. Even though we achieved an accuracy of 86%, we conducted a grid search to see if we could increase the performance even further. However, using both the machine learning models: random forest and logistic regression, the best we could get from this data was 86 percent.

Keywords: First we load and view the dataset.

I. INTRODUCTION

Credit hazard the board in banks centers around deciding whether the customer will default or her credit will break down. For instance, the benefits of the Taiwanese banks deteriorated because of the inundation of the land business, who were their significant customers. The banks then, at that point, considered stretching out their customer base to round up more advantages and in lieu of that, they started giving Mastercards and empowering an ever increasing number of people to apply for them. Over the long haul, the young people of Taiwan became their objective customers. Since the low compensation of youth, customers started defaulting on their portions and by February 2006, obligation because of Mastercards and other money cards was around 260 billion USD. This led to numerous issues in Taiwan. Self destruction rates and other criminal operations were beginning to increment to reimburse the advances from banks. Assuming forecast of the clients was done prior to giving them Mastercards dependent on specific variables, it would have been significantly more useful and proficient and would have gone far to stay away from such high obligation issues. Banks presently utilize numerous order strategies like nave bayes and KNN to investigate hazard forecast. FICO rating cards are a typical danger control strategy in the monetary business which utilize individual data of the clients and information presented by them to foresee the likelihood of future defaults and Mastercard borrowings. The bank would then be able to choose whether to give a Visa to the candidate. Financial assessments can equitably evaluate the size of hazard.

Manuscript received on November 24, 2021.
Revised Manuscript received on December 02, 2021.
Manuscript published on January 30, 2022.

* Correspondence Author
Harsha Vardhan Peela, Department of Information Technology, Manipal University, Jaipur (Rajasthan), India. Email: harsha.189302172@muj.manipal.edu
Tanuj Gupta, Department of Information Technology, Manipal University, Jaipur (Rajasthan), India. Email: tanuj.rar@gmail.com
Nishit Rathod, Department of Information Technology, Manipal University, Jaipur (Rajasthan), India. Email: nishitrathod282@gmail.com
Tushar Bose, Department of Information Technology, Manipal University, Jaipur (Rajasthan), India. Email: tushar.189302021@muj.manipal.edu
Neha Sharma*, Department of Information Technology, Manipal University, Jaipur (Rajasthan), India. Email: nehavaishnavisharma@gmail.com

II. BACKGROUND MATERIAL

A. Conceptual Overview

Banks receive a lot of credit card applications. Many of the applications do not get approved for a variety of reasons, like increased loan balances or poor-income levels. Manually analyzing these applications can be very time-consuming and full of human errors. Thankfully, we can automate this task with the help of machine learning. Below are the concepts and theories that helped understand the project solution and are an integral part of this process. A thorough understanding of them facilitated the development process. Let's first go through the different types of ML models:

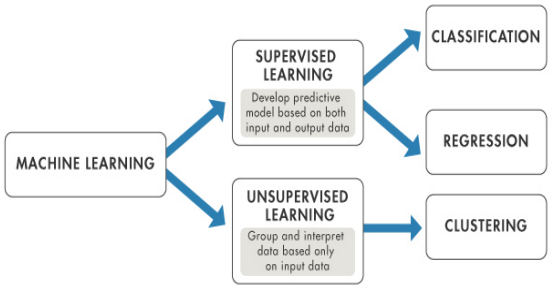

Figure 1: Types of ML Models

2.2 Supervised learning

Administered learning is use of those calculations that structure a theory dependent on

remotely provided information i.e., preparing information and afterward make future forecasts on test information. Essentially, the calculations construct a model dependent on the underlying data and afterward group the obscure information as per the indicator ascribes of the preparation information. Supervised machine learning algorithms learn by example.

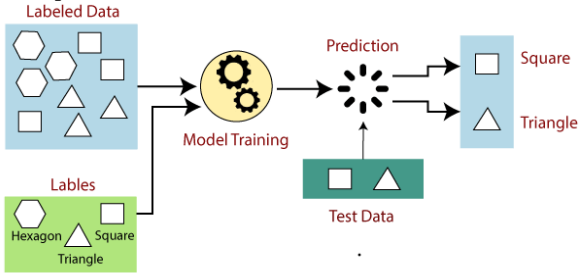

Figure 2: Supervised Learning

2.2.1 Logistic Regression

Calculated relapse is a kind of relapse investigation. Relapse examination is a sort of prescient displaying strategy which tracks down a connection between a reliant variable and possibly one autonomous variable or a progression of free factors. At the point when at least two autonomous factors are utilized to anticipate or clarify the result of the reliant variable, it is known as various relapse.

Relapse examination can be utilized for a few things:

1) Forecasting the effect of explicit changes.

2) Forecasting patterns and future qualities.

3) Determining the strength of various predictors or, all in all, evaluating the amount of an effect the free variable(s) has on a reliant variable.

Calculated Regression is prevalently utilized in credit scoring and hazard investigation techniques. It is a specific instance of a summed up direct model and the result variable, or the reliant variable is downright/paired, and the information factors can be ceaseless.

Some assumptions about logistic regression are:

1) The ward variable is double or dichotomouswhich implies that It squeezes into one of two obvious classes.

2) There ought to be no, or very little, multicollinearity between the indicator variablesin different words, the indicator factors (or the autonomous factors) ought to be free of one another. Which implies that there ought not be a high connection between's the free factors.

3) The autonomous factors ought to be straightly identified with the log chances.

4) Logistic relapse requires genuinely huge example sizesthe bigger the example size, the more solid (and amazing) one can anticipate that the results of their analysis should be.

What sorts of true situations can Logistic relapse be applied to?

- Strategic relapse can be utilized to ascertain the likelihood of a parallel occasion happening, and to take into account characterization issues. For instance:
- Foreseeing assuming that an approaching email is spam or not spam or anticipating in case a charge card ought to be supported or not.
- In a clinical setting, strategic relapse can be utilized to foresee in the event that a growth is harmful or harmless.
- In showcasing, it tends to be utilized to anticipate if a client (or gathering of clients) will purchase a specific item or not.
- A web-based training organization may utilize strategic relapse to anticipate whether or not an understudy will finish their seminar on schedule.

The logistic regression curve looks somewhat like the following graph:

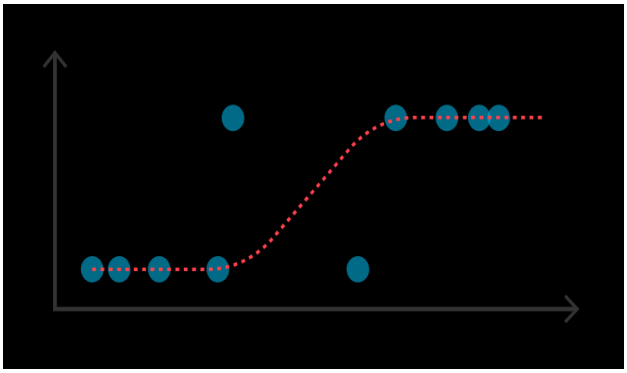

Figure 3: Logistic Regression Curve

2.2.2 Random Forest:

It is a classification algorithm. The random forest classifier works by performing row sampling and feature sampling + replacement. This means that the sum of the record or features from the dataset may get repeated. Random forest results in low variance because the data change will not impact the accuracy error rate.

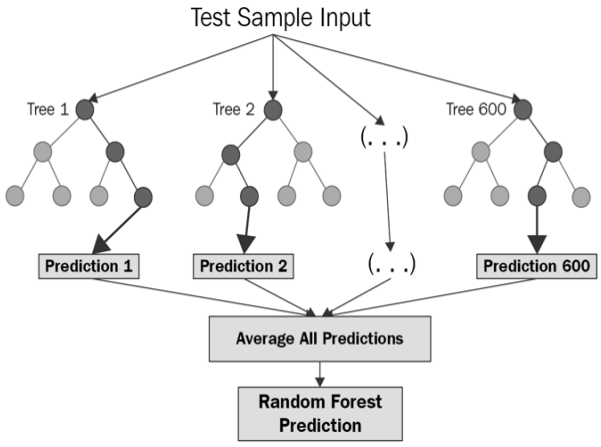

Figure 4: Random Forest

2.2.3 Label Encoding

Label Encoding means converting the labels to numeric form so that they are converted into machine-readable form. Then the machine learning algorithms will be able to decide in an efficient way as to how those labels should be operated. It is a crucial step in the pre-processing stage for the dataset in supervised learning.

2.2.4 MinMaxScaler

This converts features by rescaling them to a given range. This assessor scales and deciphers each feature separately to such an extent that it is in the given range on the training set, for example somewhere in the range of zero and one.

2.2.5 GridSearchCV

It is used in performing hyperparameter tuning in order to determine the optimal values for a given model. It finds the optimal value at which the model gives the best possible accuracy.

2.2.6 Hyperparameters:

- tol:

tol stands for tolerance (stopping criteria). It tells the scikit library to stop searching for a minimum (or maximum) once some tolerance is accomplished, i.e. once the model is close enough, tol itself will alter depending on the function that is being minimized. There is no universal tolerance to scikit.

- max_iter:

Maximum number of iterations taken for the solvers to converge.

III. METHODOLOGY

A. Inspecting the applications

The highlights in a commonplace Mastercard application are Gender, Age, Debt, Married, Bank Customer, Education Level, Ethnicity, Years Employed, Prior Default, Employed, Credit Score, Drivers License, Citizen, Zip Code, Income lastly the Approval Status. This gives us a very decent beginning stage.

B. *Handling the missing values*

As we can see from our first glance at the data, the dataset has a mixture of numerical and non-numerical features. This can be fixed with some preprocessing.

The dataset consists of both numeric and non-numeric data. The features 2, 7, 10 and 14 contain numeric values and rest of the features contain non-numeric values.

The dataset also contains values from several ranges. Some features have a value range of 0 - 28, some have a range of 2 - 67, and some have a range of 1017 - 100000. Apart from this, we can get some insightful statistical information (like mean, max, and min) about the features that have numerical values.

Finally, the dataset has missing values. These missing values in the dataset are labeled with '?', which is replaced with NaN.

C. *Pre-processing the data*

We first divide the preprocessing steps into three tasks:

- Converting the non-numeric data into numeric data.
- Split the given data into train and test sets.
- Scale the feature values to a uniform range.

First, we will convert all the non-numeric values into numeric values. This is done because not only it results in a faster computation but also many machines learning models (especially the ones developed using scikit-learn) require the data to be in a strictly numeric format.

Out[17]:

	Male	Age	Debt	Married	BankCustomer	EducationLevel	Ethnicity	YearsEmployed	PriorDefault	Employed	CreditScore	DriversLicense	Citizen	ZipC
680	b	19.50	0.290	u	g	k	v	0.290	f	f	0	f	g	
681	b	27.83	1.000	y	p	d	h	3.000	f	f	0	f	g	
682	b	17.08	3.290	u	g	i	v	0.335	f	f	0	t	g	
683	b	36.42	0.750	y	p	d	v	0.585	f	f	0	f	g	
684	b	40.58	3.290	u	g	m	v	3.500	f	f	0	t	s	
685	b	21.08	10.085	y	p	e	h	1.250	f	f	0	f	g	
686	a	22.67	0.750	u	g	c	v	2.000	f	t	2	t	g	
687	a	25.25	13.500	y	p	ff	ff	2.000	f	t	1	t	g	
688	b	17.92	0.205	u	g	aa	v	0.040	f	f	0	f	g	
689	b	35.00	3.375	u	g	c	h	8.290	f	f	0	t	g	

D. Data description and distribution

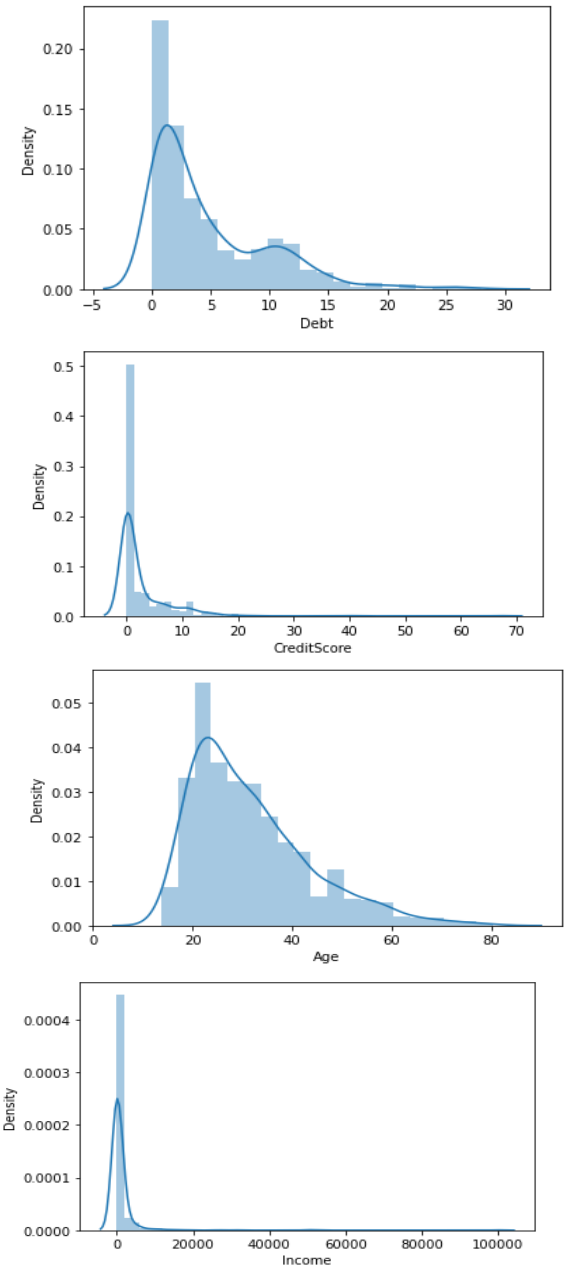

Figure 5: Density distribution plot of Age, Years Employed, Credit Score, Income

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

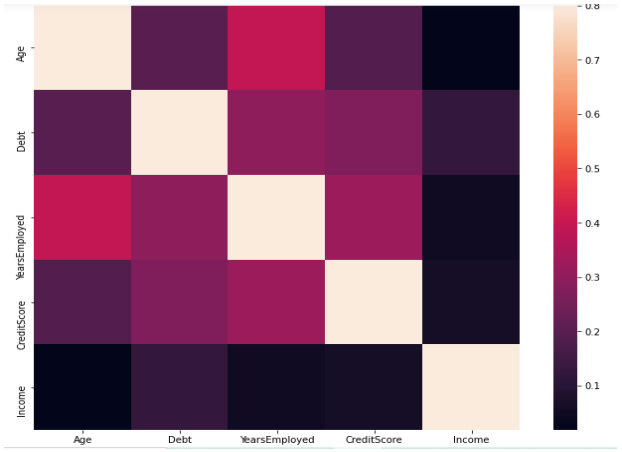

Figure 6: Heat map of correlation matrix

Figure: SNS pairplot of Age, Years Employed, Credit Score, Income

Now, split the data into a train set and test set to prepare it for two different phases of the modeling: training and testing. In ideal cases, no information from the test data is used to scale the training data or should be used to direct the training process of a machine learning model. And so, we first split the data and then apply the scaling. Also, the features: Drivers License and Zip Code are not as relevant as the other features in the dataset for predicting credit card approvals. We drop them so that our machine learning model has the best set of features. This is often referred to as feature selection.

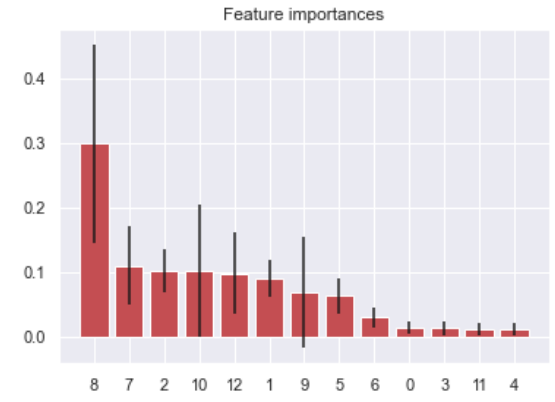

Figure 9: Features ranked on the basis of their importance

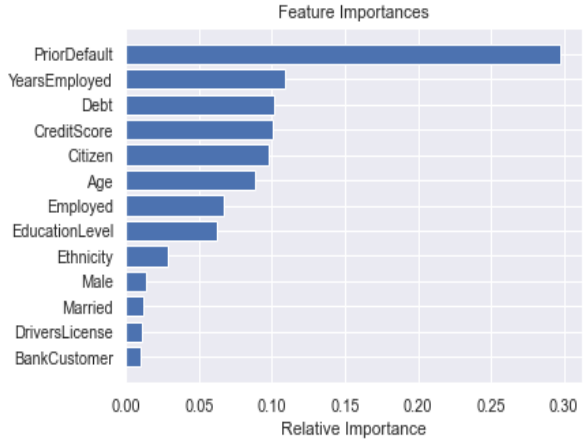

Figure 10: Features ranked on the basis of their relative importance

E. Fitting a logistic regression model to the train set

Predicting if a credit card application will be approved or not is a classification task. According to the dataset given on UCI, the dataset contains more instances that show the "Denied" status than instances of "Approved" status. Specifically, out of 690 instances, there are 383 (55.5%) applications that got denied and 307 (44.5%) applications that got approved. This gives us a benchmark. A decent machine learning model accurately predicts the status of the applications with respect to the statistics. The features that affect the credit card approval decision process are correlated with each other. Because of this correlation, we'll take advantage of the fact that generalized linear models perform well in these cases. We begin our machine learning modeling with a Logistic Regression model.

F. Tools to make the model perform better

The machine learning model was able to yield an accuracy score of almost 85%.

For the confusion matrix, the first element of the first row of the confusion matrix denotes the true negatives meaning the number of negative instances (denied applications) predicted by the model correctly. And the bottom right element of the confusion matrix denotes the true positives meaning the number of positive instances (approved applications) predicted by the model correctly.

Retrieval Number: 100.1/ijsce.B35350111222
DOI: 10.35940/ijsce.B3535.0111222
Journal Website: www.ijsce.org

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

We can perform a grid search of the model parameters to improve the model's ability to predict credit card approvals. scikit-learn's implementation of logistic regression consists of different hyperparameters but we will grid search over the following two:

- TOL
- MAX_ITER

We have defined the grid of hyperparameter values and converted them into a single dictionary format which GridSearchCV() expects as one of its parameters. Now, we begin the grid search to see which values perform best. We instantiate GridSearchCV() with our earlier logreg model with all the data we have. As an alternative to passing train and test sets separately, we will supply X (scaled version) and y. We will also instruct GridSearchCV() to perform a cross-validation of five folds.

IV. RESULTS AND ANALYSIS

Below, we analyze one by one the features that have an impact on the approval process through graphs.

1) Education level of applicants:

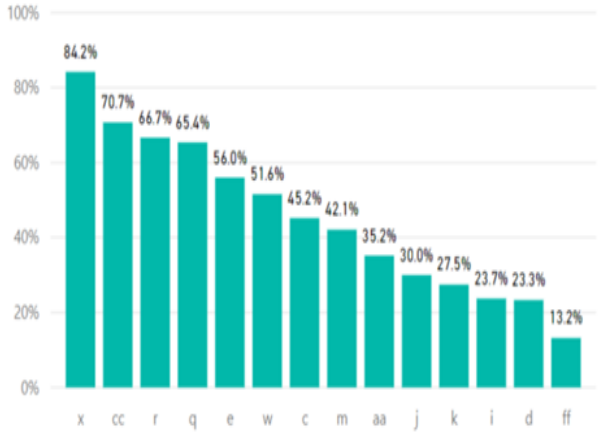

Figure 11: Education level of applicants

From the above graph of the education level, we can see that education plays a crucial role in the credit card approval process. The higher the education level, the higher the chances of getting the credit card approved i.e The person with an education level of "x" has a probability of 84% that the credit card will be approved.

2) Gender:

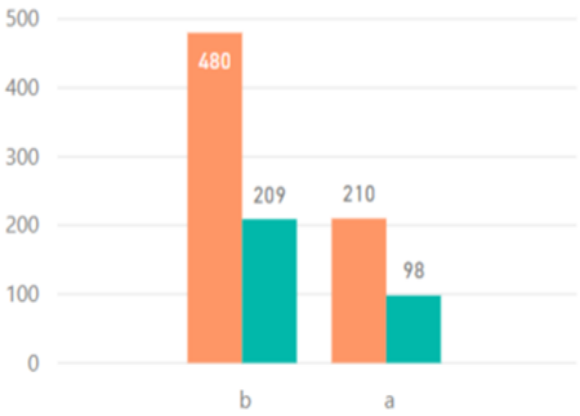

Figure 12: Gender distribution of applicants

Looking at the bar chart of the gender data, there aren't any strong connections. Out of total 690 applicants 480 are from category b and 210 from category a. However, when we look into the ratio of request approval from their respective counts, there is no big difference. It's 43.5% for a and 46.7% for b.

3) PriorDefault:

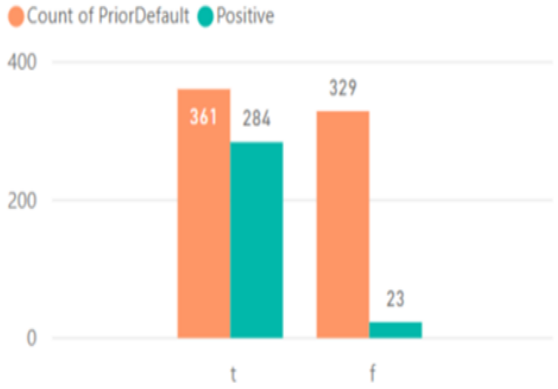

Figure 13: Count of prior default of applicants

Prior default has a major impact on the decision of credit card request approval. Out of 690, 329 don't have prior defaults whereas 361 have prior default. The ratio is balanced for both cases. But, when we look into the number of approval from both the classes it's 7% and 78.7% from their respective totals. This indicates if the applicant has no prior defaults, probability of getting the credit card request approved is much higher.

4) Credit score:

Figure 14: Distribution of Credit Score of applicants

V. CONCLUSION

As can be inferred from the above bar chart, 90% of the user's requests got approved by those who had a good credit score. And the ones with a bad credit score were unlikely of getting their credit card application approved. The feature that played the most important role in deciding whether to approve the credit card application is Prior Default. This attribute has the highest relative importance compared to all the other features. The machine learning model we built gave an 86 % accuracy for predicting whether the credit card will be approved or not, considering the various factors mentioned in the application of the credit card holder.

Even though we achieved an accuracy of 86%, we conducted a grid search to see if we could increase the performance even further. However, using both the machine learning models: random forest and logistic regression, the best we could get for this data was 86 percent.

and we can conclude that the factor that had a major impact on the decision making of whether to approve a credit card application was Prior Default.

REFERENCES

1. Koh, H. C., & Chan, K. L. G. (2002). Data mining and customer relationship marketing in the banking industry. Singapore Management Review, 24(2), 1–27.
2. S. B. Kotsiantis. Supervised Machine Learning: A Review of Classification Techniques. Informatica 31 (2007) 249-268 *Web*
3. Dataset, UCI http://archive.ics.uci.edu/ml/datasets/credit+approval
4. http://www.cs.stir.ac.uk/courses/ITNP4B/lectures/kms/4-MLP.pdf
5. http://localhost:8888/notebooks/Downloads/MajorProjectBM/Predicting%20Credit%20Card%20Approvals/notebook.ipynb
6. Medium. 2021. *Credit Card Approval Prediction Model in Python*. [online] Available at: <https://medium.com/@ashish.tripathi1207/credit-card-approval-prediction-model-in-python-c0e07677058e> [Accessed 29 June 2021].
7. http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
8. Http://rstudio-pubs-static.s3.amazonaws.com/73039_9946de135c0a49daa7a0a9eda4a67a72.html

AUTHORS PROFILE

Harsha Vardhan Peela, is currently pursuing his bachelor's in Information Technology at Manipal University Jaipur, Jaipur, India. His areas of interest include Data Science, Machine Learning, and Internet of Things**.**

Tanuj Gupta is pursuing his bachelor's in technology in Information Technology from Manipal University Jaipur. He is currently in his 4th year and has co-authored a research paper on 'Model Comparison and Multiclass Implementation Analysis on the UNSW NB15 Dataset' published in the IEEE ComPE 2021 conference. Machine Learning has been an interest of mine since the start, and I have developed fine skills in the field along the way.

Nishit Rathod is currently in 4th year pursuing bachelor's in Information Technology from Manipal University Jaipur. I am the lead author of a research paper on 'Model Comparison and Multiclass Implementation Analysis on the UNSW NB15 Dataset' published in the IEEE ComPE 2021 conference. He is certified in CEH Practical certification provided by EC-Council and currently interning at E&Y in the domain of Cyber Security.

I am Tushar Bose, and I am pursuing my bachelor's in technology focused in Information Technology from Manipal University Jaipur. I always had a keen interest in Machine Learning, and over the years I believe I possess a certain set of skills and a knack of learning which enabled me to flourish in the field of Data Science. I am currently in my fourth year of my undergrad and I have a few Machine Learning internships and projects under my belt, one of the projects being "Detection of Autism Syndrome Disorder in Children at a Young Age". Data has always fascinated me and I really believe Data Science and Machine Learning is the future.

Neha Sharma is an Assistant Professor in department of Information Technology, Manipal University Jaipur, India. She is currently pursuing her PhD. in Network Security from Manipal University Jaipur, India. She has an overall experience in industry and academics of more than 10 years. She has many National and International publications to her credit.