Fontys Hogescholen ICT

Project Proposal

Loan Prediction by Zhaklin Yanakieva

Zhaklin Yanakieva

Eindhoven, 08/03/2021

Table of content

Versioning table

Time period(weeks)	Comments on changes	Version
3 — 5	Initial changes to the document — creating the structure	1.0.0
6 — 7	Filling up with the needed information, such as explaining graphs, and summarizing	2.0.0
8 — 10	Adding the conclusion and more explanations	3.0.0

Project definition

Background

The recent significant increase in loans has generated interest in understanding the key factors predicting the non-performance of these loans.[1] The idea behind this ML project is to build a model that will classify how much loan the user can take, depending on certain data that will be required by the person, signing for a loan, such as name/education/marital status/number of dependents, employment, etc.

Project Goal

This project is going to be a prototype of a real loan predictor, which main goal is to predict whether granting the loan to a particular person will be safe or not. It will be a completely safe application for people and bank checking on the information input and deducting if the loan status is YES or NO.

Project Domain

The domain of this project is determined to be: *Prediction of a loan*

Also the following research methods were applied:

- Research
- Literature study
- Stakeholder analysis

---

[1] https://www.researchgate.net/publication/318816798_Project_Report_Student_Loan_Repayment_Prediction

Domain research

Research

What models to use after the EDA is ready?

After exploring the variety of models in Machine Learning, I decided to use the three models that I found best for these predictions: Logistic Regression, Decision tree classifier and Random forest classifier.

Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable.

Decision tree classifier:

A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

Random forest classifier:

Random forests classifier is an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification)

How to deploy the project?

For this project, I decide that the best decision for a deployment is to be a flask API, which is a web framework for Python, meaning that it provides functionality for building web applications, including managing HTTP requests and rendering templates. After that I use heroku, which is a cloud-based

development platform as a service (PaaS) provider, to put the project on a server instead of it being on localhost.

Literature study

What is required to receive a loan?[2]

Are You Aware of These Bank Loan Requirements?

1. Purpose of Loan
2. Business Experience
3. Business Plan
4. Credit History
5. Personal Information
6. Financial Statements
7. Collateral
8. Cash FlowHow to Qualify for a Personal Loan

[3]There are many steps to take to qualify for a personal loan, with the first being to make sure that it's right for you. For example, if you want to borrow money to remodel your house or buy a car, a home equity loan or an auto loan may come with a lower interest rate. Unlike unsecured personal loans based solely on your creditworthiness, these loans are secured by the home you want to fix up or the car you want to buy.

Although paying for a family vacation or consolidating debt fits into the personal loan category, you may also want to check into a 0% introductory APR credit card. If you go that route, however, be sure that you can pay off the balance before the 0% rate expires.

---

[2] https://www.forafinancial.com/blog/working-capital/8-bank-loan-requirements/
[3] https://www.investopedia.com/articles/personal-finance/010516/how-apply-personal-loan.asp

What makes people have the desire/need to have a loan?

[4]Personal loans are borrowed money that can be used for large purchases, debt consolidation, emergency expenses and much more. These loans are paid back in monthly installments over the course of typically two to six years, but it can take longer depending on your circumstances and how diligent you are with making payments.

Here are the top nine reasons to get a personal loan and when they make sense:

1. Debt consolidation.
2. Alternative to a payday loan.
3. Home remodeling.
4. Moving costs.
5. Emergency expenses.
6. Appliance purchases.
7. Vehicle financing.
8. Wedding expenses.
9. Vacation costs.

Stakeholders

Name	**Details**
Students	People who need a loan in order to finish their education and also are in need of a maintenance loan for a living.
Labour force	Working people who either need a loan for new purchases, or for living costs.

[4] https://www.bankrate.com/loans/personal-loans/top-reasons-to-apply-for-personal-loan/

Self-employed	Working for oneself as a freelance or the owner of a business rather than for an employer.
Banks	Institutions which decide whether a person can be approved for a loan or not.

Planning

In this chapter, I am going to explain the project planning for the upcoming weeks this semester and how I am going to achieve my goal. It gives me a better understanding of the structure of the semester and also helps me to deal with deadlines. I am using my own sprints that are represented in the table (see Picture 1).

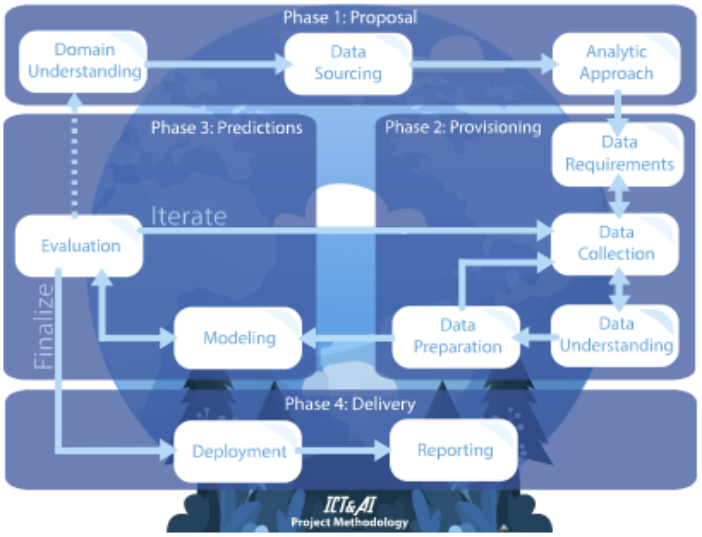
Part	Weeks	Project days	Phase(s)
A	3-6	12	1 and 2
B	7-10	10	2 and 3
C	11-12	10	4

(Picture 1:Phases)

From week 3 until week 6, this is the first phase (phase 1) of the project. This is where the business proposal and the exploratory data analysis report must be delivered, however Ialso decided to start with the first steps of the provisioning phase.

The next phase is week 7 until week 10. That is the time for the second part of the project (phase 2 and 3). The reason for doing phase 2 and 3 in the same sprint is because in the second iterations, the

product has to be improved compared to the first one. This is the same for the final part of the project (phase 2,3 and 4). That is the final sprint where I have the final product ready.



(Picture 2: Phases of the project)

As you can see on Picture 3, the 4 weeks will be spent on the Project Proposal, the EDA (Exploratory Data Analysis), Data requirement and Data collection ledger. After the 6th week I am going to spend the time on the Provisioning and Prediction phases. At week 10 to 12 , there will be the deployment of the project presented.

Step	Weeks	Phase	Deliverables
1	3 – 6	Proposal	Project Proposal, Exploratory Data Analysis Report, Data requirement, Data collection ledger
2	7 – 10	Provisioning	Data Analysis Report, Preparation notebook
3	14 – 15	Prediction, Deployment	Modelling notebook attached, Deployment report

(Picture 3: Planning of the project)

Ethical considerations

In this part of the project, we are going to use the "The Rights Approach"[5] – a method which is focused on the individual's right to choose for herself or himself. It is a violation of human dignity to use people in ways they do not freely choose, so what should be taken into account is: the right of privacy and the right to what is agreed. In these cases, applicants have the right to present certain information only to the institutions they are asking for a loan.

Ethical dilemmas:[6]

- **Is the development of "loan predictor" ethical?**

  I consider the development of this project to be ethical because it benefits people without harming anyone. Also different levels of consent required from the people assigning for a loan can be enforced.

- **What are the actions the involved stakeholders are carrying out to address the ethical concerns?**

  I assume that part of the stakeholders - students, labour force, self-employed, will not have to take the ethical concerns into account, however, the institutions that approve them for a loan will be responsible for this.

- **What consent will be processed?**

  The data that will be required so as the predictor to be manufactured will be based on the name, loan data, age, etc. In this case, the data will be processed in a way to create a result if the person is suitable or not for a loan.

- **Is there a violation of moral rights involved in the development of the "loan predictor"?**

  As soon as the privacy of the client is met, which means that only the person that is applying for a loan and the bank that gives it have the right to see the data, there is no violation.

[5] https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/thinking-ethically/
[6] https://www.nature.com/articles/s41599-020-0501-9

Impact assessment

NAME: Loan Predictor

DATE: March 8, 2021 12:46 AM

DESCRIPTION OF TECHNOLOGY

The recent significant increase in loans has generated interest in understanding the key factors predicting the non-performance of these loans. The idea behind this ML project is to build a model that will classify how much loan the user can take, depending on certain data that will be required by the person, signing for a loan, such as name/education/marital status/number of dependents, employment, etc.

TRANSPARENCY

How is it explained to the users about how a technology works and how the business model works?

I will engineer the user experience of the final product to have the main input features visible. A more detailed explanation of predictions will be placed in the 'Modelling' part of the project..

IMPACT ON SOCIETY

What is the challenge at hand? What problem (what 'pain') does this technology want to solve?

In their lives, at least once, people had the opportunity to have a loan. The prediction project will help lenders (banks) and people who want a loan, know if they are suitable for a loan. It will have a more beneficial than harmful effect to society. There can be some stigmatising effects for people who are in need for a loan, but do not cover the requirements.

HATEFUL AND CRIMINAL ACTORS

In which way can this technology be used to break the law or avoid the consequences of breaking the law?

The data that will be processed is going to be available only to the institutions that give out the loans, so in the case of creating this project, no one's privacy will be hurt.

FUTURE

What could possibly happen with this technology in the future?

As more and more people use this technology, it will become recognized how much more easily will be the conclusion if someone is approved for a loan.

PRIVACY

Does this technology register personal data? If yes, what personal data?

The technology should take the name/education/marital status/number of dependents, employment. Taking into account these labels, accurate calculations will be made to determine the loan status of a person.

HUMAN VALUES

How does your technology affect the identity of users?

The technology is rather beneficial to the users than harmful. It does happen to appear certain drawbacks for people who are unemployed or self-employed because it is usually more difficult to determine their loan status according to the data they apply. However, it is not impossible for them to be approved for a loan, so the project will be as beneficial for them as it is for the other people (students, labour force).

SUSTAINABILITY

In what way is the direct and indirect energy use of this technology taken into account?

If successful, this project will cause significant optimization in our client's energy use by encouraging them to plan their data and see if they can be approved for a loan. The product itself will be software based(web application) so that its energy consumption would be minimal and based on the user’s device.

STAKEHOLDERS

Students
Labour force
Self-employed
Banks

DATA

Are you familiar with the fundamental shortcomings and pitfalls of data and do you take this sufficiently into account in your technology?

Dependending on the available data, the research will be made, but there are always limits. This is clear to me, so the technology will be made with the awareness of these limitations and the users will be notified about them before beginning their experience with the application.

INCLUSIVITY

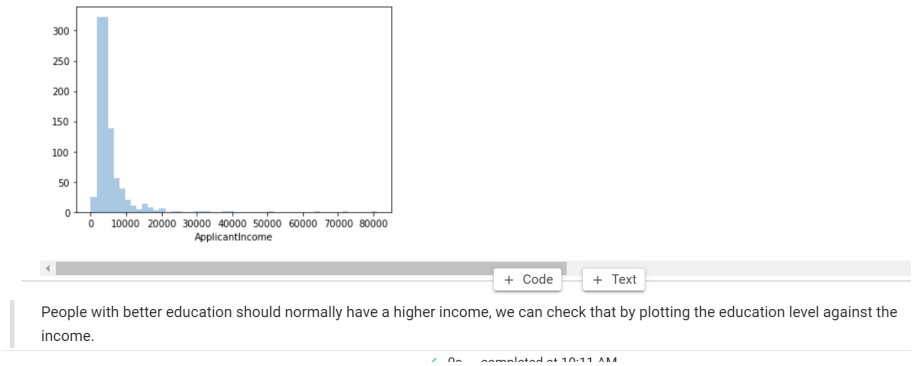
Does this technology have a built-in bias?

Due to political issues or historical issues, the banks are biased and therefore, there is some bias in the built-in of the Loan predictor.

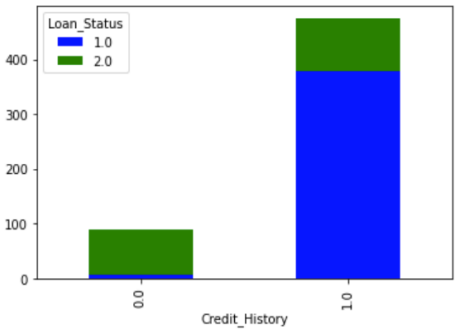
Data understanding

In the EDA, I decided to use data that is from the website 'Kaggle' where usually data sets can be found and used for analysis and creating projects. In this project, it was impossible to scrape data from a website because it is strictly personal and it would be violence of privacy. After I created a dataframe by joining two datasets, I saved it into a csv file in order to use it in the modelling phase of the project.

In the EDA, I seperate the analysis into several parts. The first part is the Data collection. For this, I decided to use extraction of the data from a csv file as a technique because it gives the opportunity to work with two data sets that are given since I cannot use any back information because of its confidentiality. Then, the following part is the Data preprocessing, for which I tried different types of data transforms to expose the data structure better, so the model accuracy may be improved later.



> Here I am exploring the distribution of the numerical variables mainly the Applicant income and the Loan amount. What can be noticed are quite a few outliers.



This shows that the chances of getting a loan are higher if the applicant has a valid credit history.

You can find the exploratory data for test analysis here or in a pdf file in the project folder..

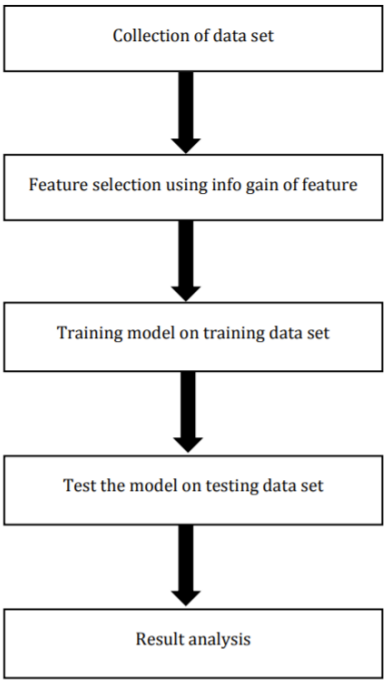
Modelling

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.[7]

The modelling stage will consist of training a Logistic Regression, Decision tree and Random Forest model. The reason for choosing to train this machine learning model is based on the explored information - the predicting label, which is a numeric value for which the models are suitable. In this case, the predicting label is the price of a real estate. The loan status is displayed as a yes-or-no answer.

[7] https://docs.microsoft.com/en-us/windows/ai/windows-ml/what-is-a-machine-learning-model

Loan prediction methodology:



Evaluation

This chapter explains how I will assess the performance of the project and what features will be optimized. After exploring the data set and training the models, a vital part of the project is the evaluation of the models performance.

First of all, the performance of the model has to be tested with common evaluation metrics. Then, an approach, which consists of predicting a future value, to evaluate the model is accurate enough to determine the performance.

I used three models to determine the accuracy - Logistic Regression, Decision Tree and Random Forest.

From the exploring of the models RMSE:

- Linear Regression score: 0.44
- Decision Tree score: 0.46
- Random forest score: 0.36

> RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. All of the models showe values in this range.

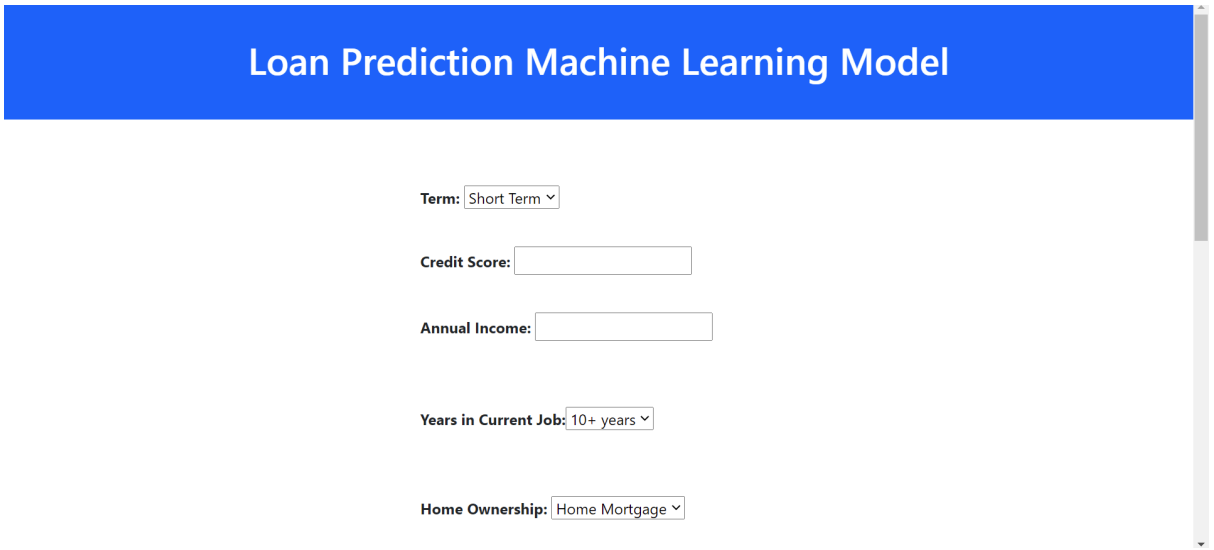
From the exploring of the models accuracy:

- Linear Regression score: 0.73 (73%)
- Decision Tree score: 0.79 (79%)
- Random forest score: 81.91 %

On the picture above is stated the information needed to compare the models and then, decide which is the best one to be chosen for the deployment phase. As a conclusion, the Random forest had the best performance of all because its accuracy stated 81% and the RMSE was a bit more than 0.36, which says that the model may well predict the score accurately.

Deployment

Ideally the model will be deployed by using an API and creating a Web application. The users will have the chance to access the Web application so as to query the API, which will be open for it.



Conclusion

The emergence of this technology has risks and ethical concerns if it is managed properly it will lead to better quality of life for human society. Be that as it may, in the event that we hit the nail on the head, we will release the full advantage of AI for mankind. The proposal demonstrates that we ought to not really believe a machine learning algorithm. It is obviously not moral to utilize a machine to make decisions when we don’t confide in it for using good ones.

Taking into account the data set, a couple of models were used in order to calculate the accuracy in the most efficient way. From the research, it could be concluded that the prototype of the project is feasible.

Tools used:

1. Numpy
2. Pandas
3. Matplotlib
4. Scikit Learn
5. Google coollaboratory
6. TCIT
7. Python

References

Forafinancial, 26th July 2019,

https://www.forafinancial.com/blog/working-capital/8-bank-loan-requirements/.

Investopedia, 22nd January,

https://www.investopedia.com/articles/personal-finance/010516/how-apply-personal-loan.asp.

Bankrate, 11th January 2021,

https://www.bankrate.com/loans/personal-loans/top-reasons-to-apply-for-personal-loan/.

Research gate, July 2017,

https://www.researchgate.net/publication/318816798_Project_Report_Student_Loan_Repayment_Prediction.

Edu, August 2015,

https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/thinking-ethically/. Accessed March 2021.

Nature, 17th June 2017, https://www.nature.com/articles/s41599-020-0501-9. Accessed March 2021.