Second Capstone Project Proposal

Project Idea 1: To predict the clients' repayment abilities. Also determine the factors or features that influence prediction by analyzing variety of alternative data--including telco and transactional information.

PROPOSAL DETAILS

1. What is the problem you want to solve?

I plan to analyze dataset provided by my Client 'Home Credit Group' on Kaggle Competition:

https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction/data

This dataset is historical loan application data to predict whether or not an applicant will be able to repay a loan. Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities. While Home Credit is currently using various statistical and machine learning methods to make these predictions, they still want help

- Better predictions of clients' repayment abilities by unlock the full potential of their data.
- Identify features that influence the prediction

The dataset contains nearly 307K records with 122 features.

SK_ID_CURR	TARGET	NAME_CONT	CODE_GEND	FLAG_OWN_	FLAG_OWN_	CNT_CHILDR	AMT_INCOM	AMT_CREDIT	AMT_ANNUI	AMT_GOODS	NAME_TYPE_	NAME_INCO	NAME_EDUC	NAME_FAMI	NAME_HOUS	REGION_POI	DAYS_BIRTH	DAYS_EMI
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.018801	-9461	-6
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher educa	Married	House / apar	0.003541	-16765	-11
100004	0	Revolving loa	M	Y	Y	0	67500	135000	6750	135000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.010032	-19046	-2
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompar	Working	Secondary /	Civil marriag	House / apar	0.008019	-19005	-30
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.028663	-19932	-30
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, part	State servant	Secondary /	Married	House / apar	0.035792	-16941	-15
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompar	Commercial	Higher educa	Married	House / apar	0.035792	-13778	-31
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompar	State servant	Higher educa	Married	House / apar	0.003122	-18850	-4
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary /	Married	House / apar	0.018634	-20099	3652

The problem that I want to solve in this project is:

- Predict whether or not an applicant will be able to repay a loan
- Determine the factors or features that influence clients' repayment abilities

2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

The client is 'Home Credit Group' and this research to predict clients' repayment abilities will be useful to them in knowing:

- What features in the dataset influence clients' repayment abilities
- External Sources of Data
- Effect of age on repayment
- Employment tenure
- Education Type

3. What data are you going to use for this? How will you acquire this data?

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

They have posted their data on Kaggle:

https://www.kaggle.com/shivamb/homecreditrisk-extensive-eda-baseline-0-772/data.

The data look like this:

SK_ID_CURR	TARGET	NAME_CONT	CODE_GEND	FLAG_OWN_	FLAG_OWN_	CNT_CHILDR	AMT_INCOM	AMT_CREDIT	AMT_ANNUI	AMT_GOODS	NAME_TYPE_	NAME_INCO	NAME_EDUC	NAME_FAMI	NAME_HOUS	REGION_POI	DAYS_BIRTH	DAYS_EMI
100002	1	Cash loans	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.018801	-9461	-6
100003	0	Cash loans	F	N	N	0	270000	1293502.5	35698.5	1129500	Family	State servant	Higher educa	Married	House / apar	0.003541	-16765	-11
100004	0	Revolving loa	M	Y	Y	0	67500	135000	6750	135000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.010032	-19046	-2
100006	0	Cash loans	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccompar	Working	Secondary /	Civil marriag	House / apar	0.008019	-19005	-30
100007	0	Cash loans	M	N	Y	0	121500	513000	21865.5	513000	Unaccompar	Working	Secondary /	Single / not r	House / apar	0.028663	-19932	-30
100008	0	Cash loans	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, part	State servant	Secondary /	Married	House / apar	0.035792	-16941	-15
100009	0	Cash loans	F	Y	Y	1	171000	1560726	41301	1395000	Unaccompar	Commercial	Higher educa	Married	House / apar	0.035792	-13778	-31
100010	0	Cash loans	M	Y	Y	0	360000	1530000	42075	1530000	Unaccompar	State servant	Higher educa	Married	House / apar	0.003122	-18850	-4
100011	0	Cash loans	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondary /	Married	House / apar	0.018634	-20099	3652

4. In brief, outline your approach to solving this problem (knowing that this might change later).

The first goal is to exhaustively determine the various relationships among Target label, Feature set and External sources of data.

1. Data wrangling will occur to clean datasets to applicable variables and check for inconsistencies in dataset as rename features, Null values/Missing values, change feature data types as required.
2. Exploratory Data Analysis (EDA) will occur to check for possible trends and/or correlations between TARGET label, other feature set attributes and external sources. The basic preliminary questions should be confidently answered after this stage.
3. Statistical inferences to conclude ‘clients’ repayment abilities’ can be predicted for any given period with the given feature set there is 95% Confidence interval.

5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The final products will include

- code ipython notebook
- Final Report
- Presentation slides

6. Is this Supervised or unsupervised problem?

This is Supervised problem as the output datasets are provided and this I used to predict the future outcomes of target variable.

7. Is it a classification or regression problem?

This is classification problem as the label is a binary variable, 0 (will repay loan on time), 1 (will have difficulty repaying loan).`

8. What variable is it that you are trying to predict?

Label ‘TARGET’ is the variable that I analyzed to find prediction with other related features in the dataset.

9. What variables will you use as predictors?

Following are independent features that may be influencing the outcome of target variable ‘ Target’ in this project:

- External Sources of Data
- Effect of age on repayment
- Employment tenure
- Education Type