MGT 6203 Group Project Proposal

TEAM INFORMATION (1 point)

Team #: 4

Team Members:

1. Keith Deuser; kdeuser3
2. Jason Duke; jduke42
3. Shravan Vudumu; svudumu3
4. Michael Daniels; mdaniels33
5. Shuxian Huang; shuang601

OBJECTIVE/PROBLEM (5 points)

Project Title: Predicting the Probability of Loan Default

Background Information on chosen project topic:

A common business scenario in the banking or financial services industry involves approving or denying loans to potential customers. Providing loans could be a very risky investment for financial institutions, so stakeholders often wish to minimize the risk of a potential customer defaulting, or not paying, their loan. Previous customer data can be used to build model(s) that can help stakeholders decide whether to approve a loan to a potential customer and minimize this risk. At the same time, the financial institutions must follow the law and not use information (like some demographic information) that might be barred from such analysis.

Problem Statement (clear and concise statement explaining purpose of your analysis and investigation):

We wish to use the loan defaulter dataset to create a predictive model(s) that will minimize the risk of approving a loan to customers with a high probability of defaulting on that loan that does not discriminate based on gender, race, religion, or other related demographics data.

State your Primary Research Question (RQ):

Which model(s) can most accurately and ethically predict the likelihood of a customer defaulting on their loan?

Add some possible Supporting Research Questions (2-4 RQs that support problem statement):

1. Can effective separate models be created by splitting the data into 3 income factors? (Low, Medium, High)
2. Not all countries follow a credit scoring system as in the US and some loans could be for international customers. Besides income and credit score, what additional variables contribute to loan default risk?
3. Are there any possible interaction variables that could be utilized to create more accurate predictions?

Business Justification: (Why is this problem interesting to solve from a business viewpoint? Try to quantify the financial, marketing or operational aspects and implications of this problem, as if you were running a company, non-profit organization, city or government that is encountering this problem.)

Solving the problem of predicting loan defaults while ensuring fairness is crucial from a business perspective due to regulatory, financial, and ethical considerations. Approving a loan always comes with a certain level of risk to a financial institution, and it is in their best interests to mitigate this risk in a manner that is compliant with local/federal laws and regulations.

DATASET/PLAN FOR DATA (4 points)

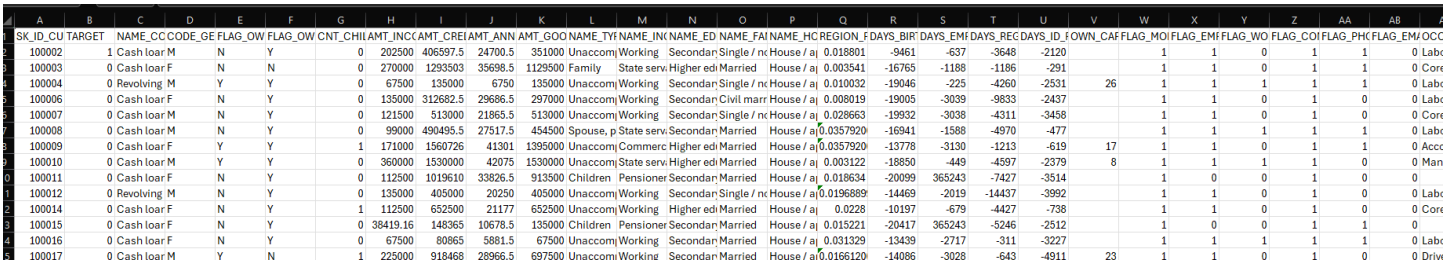
Data Sources (links, attachments, etc.):

https://www.dropbox.com/scl/fi/flia7pk7yus8cfluugwbo/home-credit-default-risk.zip?rlkey=ing5548sp8r9lh65qnolil34r&dl=0

https://www.kaggle.com/datasets/vuxminhan/home-credit-default-risk

Data Description (describe each of your data sources, include screenshots of a few rows of data):

We will be training and testing our model(s) on the application_train.csv file, which contains a response (TARGET: which can be 0 or 1), and various predictors used to classify the TARGET variable.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
1	SK_ID_CU	TARGET	NAME_CC	CODE_GE	FLAG_OW	FLAG_OW	CNT_CHIL	AMT_INCC	AMT_CRE	AMT_ANN	AMT_GOO	NAME_TYI	NAME_INC	NAME_ED	NAME_FAI	NAME_HC	REGION_F	DAYS_BIR	DAYS_EMI	DAYS_REC	DAYS_ID_I	OWN_CAF	FLAG_MO	FLAG_EMI	FLAG_WO	FLAG_COI	FLAG_PHC	FLAG_EM/	OCC
2	100002	1	Cash loar	M	N	Y	0	202500	406597.5	24700.5	351000	Unaccom	Working	Secondar	Single / nc	House / a	0.018801	-9461	-637	-3648	-2120		1	1	0	1	1	0	Labc
3	100003	0	Cash loar	F	N	N	0	270000	1293503	35698.5	1129500	Family	State serv	Higher edi	Married	House / a	0.003541	-16765	-1188	-1186	-291		1	1	0	1	1	0	Core
4	100004	0	Revolving	M	Y	Y	0	67500	135000	6750	135000	Unaccom	Working	Secondar	Single / nc	House / a	0.010032	-19046	-225	-4260	-2531	26	1	1	1	1	1	0	Labc
5	100006	0	Cash loar	F	N	Y	0	135000	312682.5	29686.5	297000	Unaccom	Working	Secondar	Civil marr	House / a	0.008019	-19005	-3039	-9833	-2437		1	1	0	1	0	0	Labc
6	100007	0	Cash loar	M	N	Y	0	121500	513000	21865.5	513000	Unaccom	Working	Secondar	Single / nc	House / a	0.028663	-19932	-3038	-4311	-3458		1	1	0	1	0	0	Core
7	100008	0	Cash loar	M	N	Y	0	99000	490495.5	27517.5	454500	Spouse, p	State serv	Secondar	Married	House / a	0.0357920	-16941	-1588	-4970	-477		1	1	1	1	1	0	Labc
8	100009	0	Cash loar	F	Y	Y	1	171000	1560726	41301	1395000	Unaccom	Commerc	Higher edi	Married	House / a	0.0357920	-13778	-3130	-1213	-619	17	1	1	0	1	1	0	Acc
9	100010	0	Cash loar	M	Y	Y	0	360000	1530000	42075	1530000	Unaccom	State serv	Higher edi	Married	House / a	0.003122	-18850	-449	-4597	-2379	8	1	1	1	1	0	0	Man
10	100011	0	Cash loar	F	N	Y	0	112500	1019610	33826.5	913500	Children	Pensioner	Secondar	Married	House / a	0.018634	-20099	365243	-7427	-3514		1	0	0	1	0	0	
11	100012	0	Revolving	M	N	Y	0	135000	405000	20250	405000	Unaccom	Working	Secondar	Single / nc	House / a	0.0196889	-14469	-2019	-14437	-3992		1	1	0	1	0	0	Labc
12	100014	0	Cash loar	F	N	Y	1	112500	652500	21177	652500	Unaccom	Working	Higher edi	Married	House / a	0.0228	-10197	-679	-4427	-738		1	1	0	1	0	0	Core
13	100015	0	Cash loar	F	N	Y	0	38419.16	148365	10678.5	135000	Children	Pensioner	Secondar	Married	House / a	0.015221	-20417	365243	-5246	-2512		1	0	0	1	1	0	
14	100016	0	Cash loar	F	N	Y	0	67500	80865	5881.5	67500	Unaccom	Working	Secondar	Married	House / a	0.031329	-13439	-2717	-311	-3227		1	1	1	1	1	0	Labc
15	100017	0	Cash loar	M	Y	N	1	225000	918468	28966.5	697500	Unaccom	Working	Secondar	Married	House / a	0.0166120	-14086	-3028	-643	-4911	23	1	1	0	1	0	0	Drive

Key Variables: (which ones will be considered independent and dependent? Are you going to create new variables? What variables do you hypothesize beforehand to be most important?)

The TARGET variable will be our dependent variable, and all other variables except for SK_ID_CURR are potential independent variables. Some of the variables we expect to be most important include AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUITY, NAME_EDUCATION_TYPE, DAYS_BIRTH, DAYS_EMPLOYED, NAME_CONTRACT_TYPE, CNT_CHILDREN, and AMT_GOODS_PRICE. Additionally, we plan on creating a dummy variable for income, split into three categories (low, middle, and high).

APPROACH/METHODOLOGY (8 points)

Planned Approach (In paragraph(s), describe the approach you will take and what are the models you will try to use? Mention any data transformations that would need to happen. How do you plan to compare your models? How do you plan to train and optimize your model hyper-parameters?))

1. Team member(s) will conduct exploratory data analysis to determine cleanliness/readiness of data and perform agreed upon transformations as needed, such as removing duplicates, correcting data entry errors, possibly creating interaction terms, identifying outliers, etc.
2. Key predictors will be identified using PCA, Logistic Regression p-values, or other significance detection methods.
3. Classification models will be used to predict the TARGET variable. Some of these possible models include:
    a. Logistic Regression
    b. SVM
    c. Regression Trees
    d. Random Forests
    e. xGBoost (Lasso)
    f. BagLearner
4. Once models have been trained and tested, we will determine the most effective model(s) by comparing the accuracy of the predicted TARGET variable.

Anticipated Conclusions/Hypothesis (what results do you expect, how will you approach lead you to determining the final conclusion of your analysis) Note: At the end of the project, you do not have to be correct or have acceptable accuracy, the purpose is to walk us through an analysis that gives the reader insight into the conclusion regarding your objective/problem statement.

We expect income, amount of credit, age, and days employed to be significant factors in classifying loan applicants. Additionally, we expect to find deeper insights into accurately predicting loan defaults by classifying customers based on their income (low, middle, or high).

What business decisions will be impacted by the results of your analysis? What could be some benefits?

The results of this model could be used to help stakeholders decide on approving or denying loans for a potential customer more accurately, which would reduce default rates among approved customers and increase customer satisfaction. Additionally, providing models that apply to different subsets of income level could help stakeholders more effectively judge loan applicants.

PROJECT TIMELINE/PLANNING (2 points)

Project Timeline/Mention key dates you hope to achieve certain milestones by:

- 3/9/2024 - Final dataset and key variables identified for analysis. Dataset is split into training and test data for training predictive models.
- 4/6/2024 - Predictive models have been trained and tested. Team decides on final predictive model(s) used to predict target values.
- 4/25/2024 – Submission of key findings and final report.