Loan Default Prediction

Abstract:

Banks provide loan to clients in exchange for the promise of repayment. Some might default on the loans; unable to repay them due to some reason. The bank maintains insurance to reduce their risk of loss in the event of default. The insured amount may cover all or just some part of the loan amount.

For this assignment, the bank wants to predict which client will default on their loans based on their financial information.

Problem Statement:

Build a logistic regression model in big data environment for a bank to predict whether a client will default on a loan or not.

Variable Description:

TARGET: Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)

- NAME_CONTRACT_TYPE: Identification if loan is cash or revolving
- CODE_GENDER: Gender of the client
- FLAG_OWN_CAR: Flag if the client owns a car
- FLAG_OWN_REALTY: Flag if client owns a house or flat
- CNT_CHILDREN: Number of children the client has
- AMT_INCOME_TOTAL: Income of the client
- AMT_CREDIT: Credit amount of the loan
- AMT_ANNUITY: Loan annuity
- NAME_INCOME_TYPE: Clients income type (businessman, working, maternity leave,)
- NAME_EDUCATION_TYPE: Level of highest education the client achieved
- NAME_FAMILY_STATUS: Family status of the client
- NAME_HOUSING_TYPE: What is the housing situation of the client (renting, living with parents, ...)
- DAYS_BIRTH: Client's age in days at the time of application
- DAYS_EMPLOYED: How many days before the application the person started current employment
- OWN_CAR_AGE: Age of client's car

• FLAG_MOBIL: Did client provide mobile phone (1=YES, 0=NO)
• FLAG_EMP_PHONE: Did client provide work phone (1=YES, 0=NO)
• FLAG_WORK_PHONE: Did client provide home phone (1=YES, 0=NO)
• FLAG_CONT_MOBILE: Was mobile phone reachable (1=YES, 0=NO)
• FLAG_PHONE: Did client provide home phone (1=YES, 0=NO)
• OCCUPATION_TYPE: What kind of occupation does the client have
• CNT_FAM_MEMBERS: How many family members does client have
• REGION_RATING_CLIENT: Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY: Our rating of the region where client lives with taking city into account (1,2,3)
• REG_REGION_NOT_LIVE_REGION: Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
• REG_REGION_NOT_WORK_REGION: Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
• ORGANIZATION_TYPE: Type of organization where client works
• FLAG_DOCUMENT_2: Did client provide document 2
• FLAG_DOCUMENT_3: Did client provide document 3
• FLAG_DOCUMENT_4: Did client provide document 4
• FLAG_DOCUMENT_5: Did client provide document 5
• FLAG_DOCUMENT_6: Did client provide document 6
• FLAG_DOCUMENT_7: Did client provide document 7
• FLAG_DOCUMENT_8: Did client provide document 8
• FLAG_DOCUMENT_9: Did client provide document 9
• FLAG_DOCUMENT_10: Did client provide document 10
• FLAG_DOCUMENT_11: Did client provide document 11
• FLAG_DOCUMENT_12: Did client provide document 12

Scope:

- Exploratory Data Analysis (EDA) using Spark SQL
- Data Pre-processing required for Spark ML
- Dealing with missing values, if any
- Training data using logistic regression

Learning Outcome:

The students will get a better understanding of how the variables are linked to each other and should be able to build a model to predict whether a client of a bank will default on a loan or not in a big data environment.