Title: Prediction of clients' repayment capabilities for an international sales finance company

Team member names: Mallik Challa, Xin Gu, Mark Johnson and Sina Kianersi

Abstract

In credit markets, people with inadequate or deficient credit history are less likely to secure a loan, though this population might be of lower income groups and perhaps in a more severe need of getting a loan. Home Credit, a Eurasian consumer finance provider, seeks to provide lending opportunities to this underserved population with less or no credit history. One main challenge is to identify clients who need loans but also are capable of repayment. In this project, we aim to predict the repayment capabilities of Home Credit clients, with a prediction diagnostic ability calculated using ROC curve. We will use the Home Credit Default Risk (HCDR) dataset that contains numerous categorical and numeric features regarding clients' loan application history, monthly credits balance, history of monthly balance of loans and credit card loans in Home Credit, and payment history. Key steps include exploratory data analysis, feature engineering, data aggregation and transformation, model selection from a variety of classification algorithms including but not limited to logistic regression, different ensembles of decision trees, Support Vector Machine, and a voting classifier. All the steps in the workflow will be done using end to end pipelines. We will report the test and training data set accuracy of the classifiers.

Data we plan to use (HDCR)

Following is a list of datasets provided. In addition to the training and test datasets, we plan to use all other datasets depending on EDA. Based on some preliminary analysis, data in 'previous_applications' and 'installments_payements' datasets appear to be key for predicting the client's repayment capabilities. However, a thorough EDA will be performed to decide on which data to include and exclude in the final model.

1. application_train.csv: Training application data at Home Credit with target (0: the loan was repaid or 1: the loan was not repaid).
2. bureau.csv: Data from other financial institutes.
3. bureau_balance.csv: Monthly data about the previous credits in bureau.
4. credit_card_balance.csv: Previous monthly credit card data.
5. installments_payments.csv: Payment history.
6. POS_CASH_balance.csv: Monthly data about previous point of sale or cash loans.
7. previous_application.csv: Previous applications of clients.

Which machine learning algorithm are you considering using and why?

Given the huge volume of data and a high number of input features it is worth looking at dimension reduction algorithm like PCA and also use SelectKbest algorithm for feature selection.

Based on EDA, there is a good possibility that some of the features are likely to be highly correlated, and hence consolidating those down will help with efficiency of the final model while still maintaining much of their information.

Since the data is labeled and the target variable is binary, supervised classification algorithms will be explored. Logistic Regression which is a simple but very powerful algorithm used commonly in practice with excellent results could be a very logical place to start when dealing with a categorical problem and will be used as the baseline model.

Other classification algorithms that will be considered include Decision Trees via random forest and gradient boosting, Support Vector machine learning and Stochastic Gradient Decent classifier.

K-Nearest Neighbor may also be considered, but with the sheer number of features present, could potentially be too computationally expensive to be worth investing in, considering its relative simplicity. Experimentation with subsets will bear all this out.

All algorithms will be tuned and optimized with a variety of hyperparameters.

Metrics that you might use to measure success (standard metrics and domain specific metrics)?

We will use area under the ROC curve to assess the prediction accuracy of our classifiers. Further, we will evaluate the models with accuracy score.

Description of the pipeline steps you plan to use

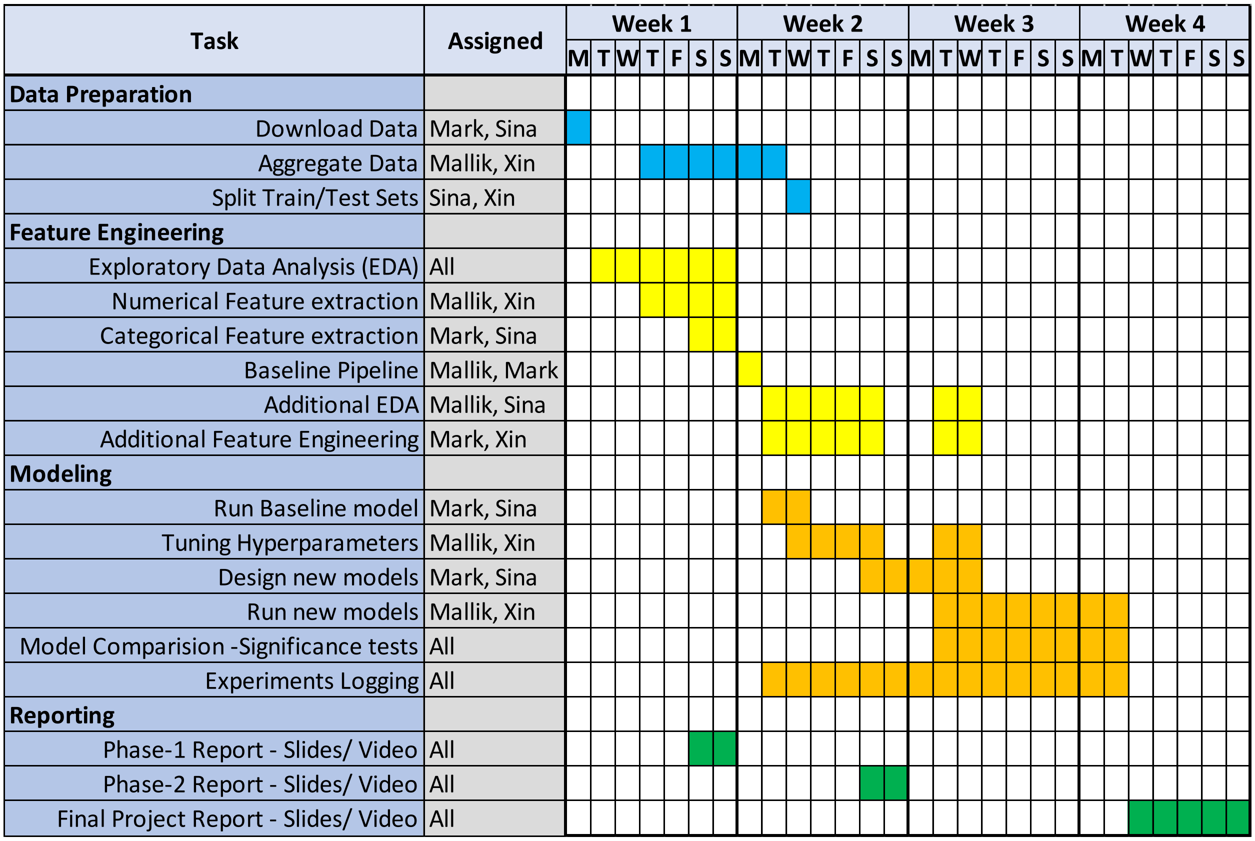
The data consists of a mix of numerical features and categorical features. However, there are no columns with ‘text’ data. Based on an initial inspection of the data, we can see that there are missing values in a significant number of columns and will require imputing. Also, combining features from different datasets will require custom transformations.

Following pipelines will be used in the workflow:

- **Numerical pipeline** - This pipeline will include but not limited to scaling of numerical features using StandardScaler() and imputing missing values using SimpleImputer().
- **Categorial pipeline** - This pipeline will include but not limited to encoding categorical features using OneHotEncoding() and imputing missing values using SimpleImputer().
- **Pre-Processing pipeline** - A prep-processing step will combine the numerical and categorical pipeline and also include steps to transform different columns, feature engineering wo add new features from different datasets. The pipeline will be built using ColumnTransformer() and will include custom transformation steps which can be fed into the pipeline directly.
- **Model pipeline** - Finally the model pipeline will combine above created pre-processing pipeline and the actual algorithm being used for classification.

Data will be fitted into the final model pipeline and transformed accordingly and used for evaluating the different models.

Key steps involved in completing this task and timeline and tentative task owners:



Task	Assigned	Week 1							Week 2							Week 3							Week 4						
		M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S
Data Preparation																													
Download Data	Mark, Sina	🟦																											
Aggregate Data	Mallik, Xin				🟦	🟦	🟦	🟦	🟦	🟦																			
Split Train/Test Sets	Sina, Xin										🟦																		
Feature Engineering																													
Exploratory Data Analysis (EDA)	All		🟨	🟨	🟨	🟨	🟨	🟨																					
Numerical Feature extraction	Mallik, Xin				🟨	🟨	🟨	🟨																					
Categorical Feature extraction	Mark, Sina						🟨	🟨																					
Baseline Pipeline	Mallik, Mark								🟨																				
Additional EDA	Mallik, Sina									🟨	🟨	🟨	🟨				🟨	🟨											
Additional Feature Engineering	Mark, Xin									🟨	🟨	🟨	🟨				🟨	🟨											
Modeling																													
Run Baseline model	Mark, Sina									🟧	🟧																		
Tuning Hyperparameters	Mallik, Xin										🟧	🟧	🟧				🟧	🟧											
Design new models	Mark, Sina													🟧	🟧	🟧	🟧	🟧											
Run new models	Mallik, Xin																🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧					
Model Comparision -Significance tests	All																🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧					
Experiments Logging	All									🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧					
Reporting																													
Phase-1 Report - Slides/ Video	All						🟩	🟩																					
Phase-2 Report - Slides/ Video	All													🟩	🟩														
Final Project Report - Slides/ Video	All																								🟩	🟩	🟩	🟩	🟩