Project Plan

Loan Sharks

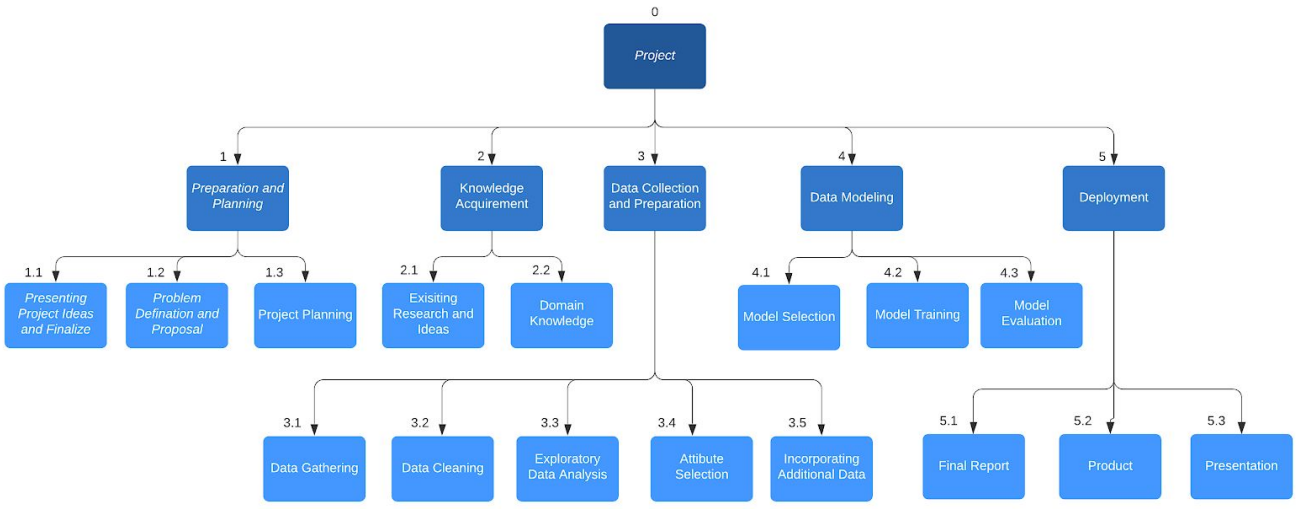
Overview

Our project aims at analyzing conventional home loans from the HMDA dataset and build a model that predicts loan approval for an applicant after removing inherent bias, if any, in the dataset. We would like to use Classification techniques such as Logistic Regression and Decision trees for the same and compare their performance to pick the one that works best for our problem definition. Doing so will help us determine what parameters can be eliminated to ensure that applicants have a fair chance at securing a loan solely on the basis of their financial standings.

Deliverables List

We plan to use some project management knowledge to arrange our work. We will make work breakdown structure in the format of responsibility assignment matrix to demonstrate our deliverables list.

Work Breakdown Structure (WBS)



Responsibility Assignment Matrix (RAM)

P : Primary Person Responsible

Item No.	WBS Item	PROJECT TASK	DETAIL OF PROJECT TASK	Ananta	Omkar	Bhuvnesh	Virat	Xiaoman
1	1.1.1	Presenting Project Idea and Finalization	Discussion on project ideas		P			
	1.1.2		Search for suitable datasets				P	
	1.1.3		Finalize problem definition					P
	1.1.4		Finalize Dataset			P		
	1.1.5		Finalize project deliverables				P	
	1.2.1	Problem Definition and Proposal	Draft Preliminary proposal			P		
	1.2.2		Refine Preliminary proposal			P		
	1.2.3		Submit preliminary project proposals		P			
	1.3.1	Project Planning	Project Plan discussion		P			
	1.3.2		Basic description of deliverables			P		
	1.3.3		Work breakdown structure				P	
	1.3.4		Responsibility Assignment Matrix				P	
	1.3.5		Draft Project Plan in a document					P
	1.3.6		Asana Management	P	P	P	P	P

		Presentation Slides	Incorporate into Presentation slides					
2	2.1.1	Exiting Research and Ideas	Look for some academic research papers relating to our topic and draft a short summary on a paper that you think is most related to our topic.	P				
	2.1.2		Look for some topics related to loans like auto insurance		P			
	2.2.1	Domain Knowledge	Research on factors responsible to adjudicate the home loan					P
	2.2.2		Research on how home loan applications are processed on US				P	
3	3.1.1	Data Gathering	Download HMDA dataset for target states	P				
	3.1.2		Download Census dataset or access it through census api			P		
	3.2.1	Data Cleaning	Remove unnecessary columns					P
	3.2.2		Resolve problems with data - Missing value, Invalid Values			P		
	3.2.3		Resolve problems with data - Data Range, Units, Duplicate Entries, Duplicate factors				P	
	3.2.4		Resolve problems with data- Linear Dependency, Changing meaning values		P			
	3.3.1	Exploratory Data Analysis	Perform Univariate Visualization- Summary statistics, CDF, PDF			P		

	3.3.2		Perform Bi-variate Visualization- Understand relationship - Boxplots, Violin plots					P
	3.3.3		Perform Multivariate Visualization- Understand interactions - Pair plots, 3-D scatter plots		P			
	3.3.4		Transform and visualize analysis for target states	P				
	3.3.5		Summarize findings				P	
	3.4.1	Attribute Selection	Select features either using Univariate selection or Correlation matrix with heat map		P			
	3.4.2		Select features which reduce Overfitting, improve Accuracy and reduce training Time			P		
	3.5.1	Incorporating Additional Data	Incorporate Additional Data			P		
		Presentation Slides	Incorporate into Presentation slides					
4	4.1.1	Model Selection	Explore appropriate algorithms to create model			P		
	4.1.2		Select few models relevant to the problem(Other than Logistic Regression and Decision Trees)					P
	4.2.1	Model Training	Split the data into training and validation sets				P	
	4.2.2		Train a logistic regression model using "loan application decision" as target variable		P			

	4.2.3		Train a decision tree model using "loan application decision" as target variable				P	
	4.2.4		Train other selected model/s using "loan application decision" as target variable	P				
	4.3.1	Model Evaluation	Evaluate the performance of Model according to the defined metrics		P			
	4.3.2		Figure out the model that gives best results.					P
	4.3.3		Testing and Improvement	P				
		Presentation Slides	Incorporate into Presentation slides					
5	5.1.1	Report	Model Evaluation					P
	5.1.2		Plot Organization			P		
	5.1.3		Code Organization	P				
	5.1.4		Article Combination / Citation/ References					P
	5.1.5		Peer evaluation form				P	
	5.1.6		Submission of Report		P			
	5.2.1	Product	Final push on Git		P			
	5.2.2		Demo of the model			P		
	5.2.3		Submission of Code	P				
	5.3.1	Presentation	Making Slides			P		
	5.3.2		Practice Presentation	P	P	P	P	P

	5.3.3		Attend Presentation	P	P	P	P	P

Major Models

Model No.1: Logistic Regression

We will build a model based on the location, gender, race, financial standing of the applicant and loan category. The logistic regression model will allow us to predict the probability with which a certain candidate's loan application will be accepted or not. It will help us determine the parameters that are inducing bias in the dataset and help us eliminate them if possible.

Model No.2: Decision Trees

Another model we will use is Decision Trees. We will be comparing the performances of both the models since they give us the same results but differ in approach. This model will reduce the space into smaller regions whereas the Logistic Regression approach will divide it using a line (planes for highly dimensional data). After building both the models we can cross validate and confirm the better approach. Finally we will visualize the results to gain a better understanding and use it to draw conclusions.

Data Source

Data Overview

This HMDA or home mortgage disclosure dataset provides the data if the applicant got the loan approved for a house purchase. There are many outcomes of the dependent variable. The loan is either originated or not passed. There are many reasons for the denial of loan, including that the consumer was denied, didn't complete the application, or something else happened. The data also have information about pre-approvals and loans sold from one institution to another.

Data Description

HMDA Dataset.

We have available data for various states in the United States and for different years starting from 2007 to 2017. We will analyze the data for different locations in a single state such as Illinois for multiple years. To ensure that we do not have data only from durations when there was a housing boom or slump we have taken into consideration the recession years 2008 and 2009 and the boom years 2011, 2012 and 2017.

Hmda_2017_il_all-records_labels.csv : This is one of the tables present in our HMDA dataset. It has columns loan type name, loan purpose name, applicant ethnicity name, applicant race name, applicant income, action taken to name a few. Other tables in the dataset have the same column names and hence have not been mentioned. Each table provides the number of observations for a given state (Illinois in this case) for a given year (2017).

We might incorporate additional data such as the Census Data and Unemployment data for a more accurate analysis. The links for the same have been given below :

BLS unemployment dataset - https://www.bls.gov/lau/#tables

Census dataset - https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

KPI Tools

Asana: Each task in the list will be accomplished on time by the appointed member and we should make sure the whole process is going within schedule.

Bitbucket: Team members have to submit their efforts on each task in bitbucket to let the project be traceable and make it more convenient to share each single idea and avoid the conflict.