Capstone Proposal

Domain Background
In the UK total household debt has increased 7% in the past 5 years with a 19% increase in unsecured credit alone. However, council tax arrears have also increased 12% and water bill arrears have increased 17% in that same period which is an 'indicator of financial distress' according to theguardian.com
This suggests that the current rate of lending is unsustainable and lending facilities that still need to be able to offer credit products to generate revenue need to make very smart decisions.

Over the next year I plan to pay off debt and improve my credit score enough to be approved for a mortgage. My plan to achieve this has involved researching patterns of behaviour that may increase my credit profile. I am interested in solving for this problem and gaining some insight into the most important factors that are used when making lending decisions.

Previous research has been undertaken in this area in particular applying machine learning to consumer credit risk models by MIT in the 2010 draft paper 'Consumer Credit Risk Models via Machine-Learning Algorithms'.

Problem Statement
The risk of lending credit remains high when limited information is used to make lending decisions. There are many factors that may affect the likelihood that someone will be unable to meet the requirements of their credit agreement and fall behind on payments.

There is a requirement for a model that can accurately predict the suitability of a person for credit, based on multiple factors from their credit file.
Machine learning can be applied to identify the importance of each of the available data attributes in making lending decisions and predict the probability that lending is suitable for that person. The columns available to the model should be variable allowing for any cleaned and restructured set of data to be used in it.

Datasets and Inputs
The dataset is available to people who enter the Home Credit Default Risk Kaggle competition.
The datasets can be linked on a single column to build one row of data for each potential lender. The shape of the data supplied is described in the table below. All of the csv files are expected to be used but not all of the columns from each csv file will be used.
This data is provided by home Credit and so should be appropriate for solving this problem.

Data Source	***Description***	***Columns***	***Rows***
application_test.csv	*Main table with test data*	*121*	*~48,700*
application_train.csv	*Main table with training data*	*122*	*~308,000*
bureau.csv	*List of clients previous credits*	*17*	*~1,720,000*
Bureau_balance.csv	*Monthly balance of previous credits*	*3*	*~27,300,000*
credit_card_balance.csv	*Monthly HC credit card balance*	*23*	*~3,840,000*
HomeCredit_columns_description.csv	*Descriptions for columns in these data files*	*5*	*219*
Instalments_payments.csv	*History of repayments on HC loans*	*8*	*~13,600,000*
POS_CASH_balance.csv	*Loan balance snapshots*	*8*	*~10,000,000*
Previous_applications.csv	*Previous application for HC credit*	*37*	*~1,670,000*
Sample_submission.csv	*Sample file with customer and probability*	*2*	*~48,700*

Solution Statement
A model will be created that will output the weights of each column used and a probability between 0-1 of the likelihood of that client being suitable for credit.

Benchmark Model

A linear regression model that accepts cleaned data predicts randomly would be the model to beat. An accuracy score of below 0.5 would be expected for this model.

Evaluation Metrics
The evaluation metric for this project will be the accuracy score of a model predicting the probability that a client will be suitable for being accepted for credit.
As this is a current Kaggle competition the score is evaluated on area under the ROC curve. I will use the accuracy score of the model e.g GridSearch.score and on area under the ROC curve.

Project Design

I plan to use a grid search algorithm to produce a model that can predict suitability of a client for lending and I will adjust the hyperparameters and regression models that are used by Grid search to improve the model.

I will begin with a data exploration stage where I will link the data sets and perform some visual representation of the data.

I will then plan the initial model and algorithm, picking initial hyperparameters and the classifiers I want to use. I may run the data set through multiple models and pick just a few to train with Grid Search.

I will clean the data and look for columns that need normalisation or one hot encoding. I will decide if I need to remove any rows of data if there are outliers that I cannot fix.

I will run Grid Search with the same hyperparameters for the chosen classifiers and then when I've set initial benchmarks for each of the classifiers I will start refining the hyperparameters of those classifiers.
I may decide the Grid Search is not the appropriate classifier to use.

Hopefully one of the regression models will prove to have the most promise in achieving a decent accuracy score and then I can concentrate on fine tuning that model.

Resources
https://www.theguardian.com/business/2017/sep/18/uk-debt-crisis-credit-cards-car-loans
https://www.kaggle.com/c/home-credit-default-risk/kernels
http://mitsloan.mit.edu/media/Lo_ConsumerCreditRiskModels.pdf