**DATA REPORT**

**INTRODUCTION**

**BACKGROUND INFORMATION**

SuperLender is a local digital lending company, which prides itself in its effective use of credit risk models to deliver profitable and high-impact loan alternatives. Its assessment approach is based on two main risk drivers of loan default prediction:

1) willingness to pay and

2) ability to pay.

Since not all customers payback, the company invests inexperienced data scientists to build robust models to effectively predict the odds of repayment.

These two fundamental drivers need to be determined at the point of each application to allow the credit grantor to make a calculated decision based on repayment odds, which in turn determines if an applicant should get a loan and if so - what the size, price and tenure of the offer will be.

There are two types of risk models in general:

**New business risk**, which would be used to assess the risk of application(s) associated with the first loan that he/she applies. The second is a **repeat or behaviour risk model**, in which case the customer has been a client and applies for a repeat loan. In the latter case - we will have an additional performance on how he/she repaid their prior loans, which we can incorporate into our risk model.

**OBJECTIVES**

Build a loan default prediction model, with a binary outcome variable where 1 is a good loan in that the customer is most likely to repay the loan. and 0 is a bad loan in that the customer is highly likely to default on repayment.

**PROJECT PLAN**

The overview plan for the study is as follows**:**

|  |  |  |
| --- | --- | --- |
| **Phase** | **Time** | **Resources** |
| Problem definition, Objectives and goals | 6 hours | Research and report writing team |
| Data sourcing and understanding | 6 hours | Data cleaning team |
| Data preparation and cleaning | 6 hours | Data cleaning team |
| Feature Engineering | 6 hours | Analytics team |
| Analysis | 12 hours | Analytics team |
| Hypothesis Testing | 4 Hours | Analytics team |
| Modelling | 5 Hours | Analytics team |
| Evaluation | 3 Hours | All teams |
| Recommendation | 12 hours | Analytics team |
| Next steps | 1 week | All teams |

**DATA SOURCING**

Data sourcing involves accessing the data and exploring it using tables and graphics that can be organized to enable you to determine the quality of the data and easily describe the results of these steps. This study used data from [Zindi](https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data) which works with organisations to determine which datasets are needed to solve a challenge by assisting in building the datasets in the appropriate format.

**DATA PREPARATION AND QUALITY**

The research identifies the procedures and techniques that were used in the collection, processing and analysis of data as well as picking the most efficient and effective libraries for computations.

We noticed that the project dataset had three tables. These tables being related to each other by the primary key (customerid).

**RESEARCH DESIGN**

Descriptive studies are usually the best methods for collecting information that will demonstrate relationships and describe the world as it exists. These types of studies are often done before an experiment to know what specific things to manipulate and include in an experiment. Descriptive studies can answer questions such as “what is” or “what was.” Experiments can typically answer “why” or “how.” The focus of this study was to establish the relationships between variables of interest and not the causal effects. It is important to note that just because variables are related, does not necessarily mean that one directly causes the other. This study was descriptive in nature and involved quantitative analysis of data.

**DATA CLEANING**

Import the libraries to be used i.e pandas, NumPy, seaborn and matplotlib.pyplot. Then load our dataset file into the environment.

Isolating columns of interest

Renaming the columns to previous names and removing trailing spaces

Merging the various datasets we had for both train and test ie:

* [**Traindemographics.csv**](https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data)
* [**Trainperf.csv**](https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data)
* [**Trainprevloans.csv**](https://zindi.africa/competitions/data-science-nigeria-challenge-1-loan-default-prediction/data)

**FEATURE ENGINEERING**

In machine learning, feature engineering is the process of extracting features from raw data using simple techniques or more advanced data mining techniques. Our dataset comprises some columns that can be toyed around with to create new features with some specific characteristics. This is important because the prediction model algorithms require such features to work properly. Our main goals in feature engineering are;

* Preparing the raw dataset and making it compatible with our algorithms in prediction modelling
* To improve the overall performance of our model

New features extracted from our raw dataset were;

* **age** - It was obtained subtracting the current date from the birthdate

Tdemographics = Tdemographics[~Tdemographics.customerid.duplicated()]

Tdemographics['age']= round((dt.date.today() - Tdemographics.birthdate.dt.date).dt.days/365, 0)

Tdemographics.drop('set', axis='columns', inplace=True)

* **loanapproval** - This is the time taken for loan approval. It was obtained getting the difference of the creation and approval dates.

Tprevious['loanapproval'] = Tprevious['approveddate'] - Tprevious['creationdate']

* **num\_loans** - This is the number of loans a customer has taken. It was obtained by counting the number of system loan ids for each client's customer id.

num\_loans = Tprevious.groupby('customerid')['systemloanid'].count().to\_frame().reset\_index()

num\_loans.rename(columns = {'systemloanid': 'num\_loans'}, inplace=True)

Tprevious = pd.merge(Tprevious, num\_loans, how='left', on='customerid')

* **average\_loanamount** - This is the average amount of total loans taken by the client.

averageloanamount = Tprevious.groupby('customerid')['loanamount'].mean().to\_frame().reset\_index()

averageloanamount.rename(columns={'loanamount': 'average\_loanamount'}, inplace=True)

Tprevious = pd.merge(Tprevious, averageloanamount, how='left', on='customerid')

* **average\_termdays** - This is the average term days for the loan between 30 to 90 days. It was obtained by getting the average of the total number of term days of the total number of loans taken by the client.

average\_termdays = Tprevious.groupby('customerid')['termdays'].mean().to\_frame().reset\_index()

average\_termdays.rename(columns={'termdays':'average\_termdays'}, inplace=True)

Tprevious = pd.merge(Tprevious, average\_termdays, how='left', on='customerid')

* **average\_diff**  - This is average days of earliness or lateness by a client to repay a loan. It was obtained by getting the difference between the first repaid date and the due date, then getting the average of this value for all the loans a customer had taken.

Tprevious['firstrepaiddate'] = Tprevious['firstrepaiddate'].astype('datetime64[ns]')

Tprevious['diff'] = (Tprevious['firstrepaiddate'] - Tprevious['firstduedate']).dt.days

Followed by:

average\_diff = Tprevious.groupby('customerid')['diff'].mean().to\_frame().reset\_index()

average\_diff.rename(columns={'diff':'average\_diff'}, inplace=True)

Tprevious = pd.merge(Tprevious, average\_diff, how='left', on='customerid')

* **average\_totaldue** - This is the average total dues by the client for all the loans he/she had taken. It was obtained by getting the average of the total dues for all the loans a client had taken.

average\_totaldue = Tprevious.groupby('customerid')['totaldue'].mean().to\_frame().reset\_index()

average\_totaldue.rename(columns={'totaldue':'average\_totaldue'}, inplace=True)

Tprevious = pd.merge(Tprevious, average\_totaldue, how='left', on='customerid')

The Tprevious was merged with Tdemographics dataset to create a new dataset demographics\_previous which was then merged with TPerformance to create our final\_df dataframe.

The final cleaned and prepared dataset was then loaded to a new csv file to be used for analysis.

**ANALYSIS**

The data was analyzed using python and the libraries( pandas and numpy). The notebook containing the analysis for this project as well as the data sets are accessible through the following link to the GitHub repository:

**Steps taken during analysis**

Perform a descriptive statistics

Feature engineering

Perform univariate and bivariate analysis

Hypothesis testing

**Finding on the analysis**

Most people that borrow money are permanently employed. Only 3 of those that borrowed loans are on contract.

GT bank has the highest number of clients that borrow money, while Unity bank has the lowest.

People between the ages of 30-40 years have the highest borrowing rate

22% of the clients are defaulters. This is higher than the global default rate which was 2.1% in 2018.

It takes around 1 hour before a loan is approved. However, the highest time is 38hours that is around 1 day and 14hours which is quite slow.

The mean age to borrow money is 36years. This is those who are working and have taken a lot of responsibilities in the past 10 years of their life.

The average defaulting rate is 2 days which is moderate. Though every lender's goal is to have a defaulting rate of 0 days as defaulting increases the risk of bad debt

More people borrow an average of 10,000 shillings worth of loans.

Those with fewer loans have a 50-50 chance of paying their loans

Customers seem to borrow the most in June and the lowest in July. This might be because it's the mid-year. After mid-year, the rate of borrowing significantly decreases

Most people that borrow their own savings accounts. This is a good thing as savings act as securities to most clients.

**HYPOTHESIS TESTING**

Several hypotheses tests were carried out in our model. We used a Chi-Square Test to compare categorical features and a Z-Score to compare categorical and numerical features. The P-Value was obtained from both categories of tests to help us in concluding the null hypotheses stated. A hypothesis\_df data frame was created for use in hypothesis testing. Using an online sample size calculator, we keyed in the population size and a significance level of 5% to obtain the sample size of 354.

* Population size: 4368
* The margin of error: 5%
* Confidence Level: 95%
* Sample size: 354

**Test 1: Does the age of a customer influence their default behaviour on loans?**

1. H0: Age of customer has no association with their loan default behaviour

2. H1: Age of a customer has an association with their loan default behaviour (Claim)

Critical value: 5.991464547107979, test\_statistic: 3.535162450352401, alpha: 0.050000000000000044, p\_value: 0.17074548463657646

* **Interpretation:** p-value is greater than alpha thus lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**Test 2: Does the number of loans that a customer has taken influence their loan default behaviour?**

1. H0 = Average number of loans is equal for both good and bad loans

2. H1 = Average number of loans != for good and bad loans (Claim)

Z statistic: 2.587012732760006, p\_value: 0.00968120146473318, alpha: 0.05

* **Interpretation:** p-value is less than alpha thus significant evidence to reject the null hypothesis with 95% confidence.
* There's sufficient statistical evidence to support the claim.

**Test 3: Does the average time taken for a customer to pay back a loan influence their loan default behaviour?**

1. H0: Average number of days to pay back a loan is equal for both good and bad loans.
2. H1: Average number of days to pay back a loan is not equal for both good and bad loans.

Z statistic: -6.448168270256521, p\_value: 1.1321003334493063e-10, alpha: 0.05

* **Interpretation:** p-value is greater than the alpha thus there is a lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**Test 4: Does the average time taken for a customer's loan to get approved influence their loan default behaviour?**

* H0: Average time to get a loan approved is equal for both good and bad loans.
* H1: Average time to get a loan approved is not equal for both good and bad loans (Claim)

Z statistic: -2.3885700378353825, p\_value: 0.01691408409591196, alpha: 0.05

* **Interpretation:** p-value is greater than the alpha thus there is a lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**Test 5: Does the bank that a customer gets a loan from influence their loan default behaviour?**

* H0: There is no association between the bank name and the good\_bad\_flag variables
* H1: There is an association between the bank name and the good\_bad\_flag variables (Claim)

Critical value: 26.29622760486423, test\_statistic: 19.84023065582578, alpha: 0.050000000000000044, p\_value: 0.22750296825965244

* **Interpretation:** p-value is greater than the alpha thus there is a lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**Test 6: Does the bank account type that a customer has influenced their loan default behaviour?**

1. H0: There is no association between the bank account type and the good\_bad\_flag variables
2. H1: There is an association between the bank account type and the good\_bad\_flag variables

Critical value: 5.991464547107979, test\_statistic: 2.259884201983607, alpha: 0.050000000000000044, p\_value: 0.3230519602688761

* **Interpretation:** p-value is greater than the alpha thus there is a lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**Test 7: Does the employment status of a customer influence their loan default behaviour?**

1. H0: There is no association between the employment status and the good\_bad\_flag variables
2. H1: There is an association between the employment status and the good\_bad\_flag variables (Claim)

Critical value: 11.070497693516351, test\_statistic: 6.195010875170066, alpha: 0.050000000000000044, p\_value: 0.28770339138551476

* **Interpretation:** p-value is greater than the alpha thus there is a lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**Test 8: Does whether a customer is referred to or not influence their default behaviour?**

1. H0: No association between referred by and good\_bad\_flag variables
2. H1: There is an association between referred by and good\_bad\_flag variables (Claim)

Critical value: 3.841458820694124, test\_statistic: 0.11800482045835192, alpha: 0.050000000000000044, p\_value: 0.7312085782483395

* **Interpretation:** p-value is greater than the alpha thus there is a lack of significant evidence, we fail to reject the null hypothesis with 95% confidence.
* There's no sufficient statistical evidence to support the claim.

**MODELLING**

The process of modelling means training a machine-learning algorithm to predict the labels from the features, tuning it for the business need, and validating it on holdout data. The output from modelling is a trained model that can be used for inference, making predictions on new data points.

In the modelling process, the following steps were carried out:

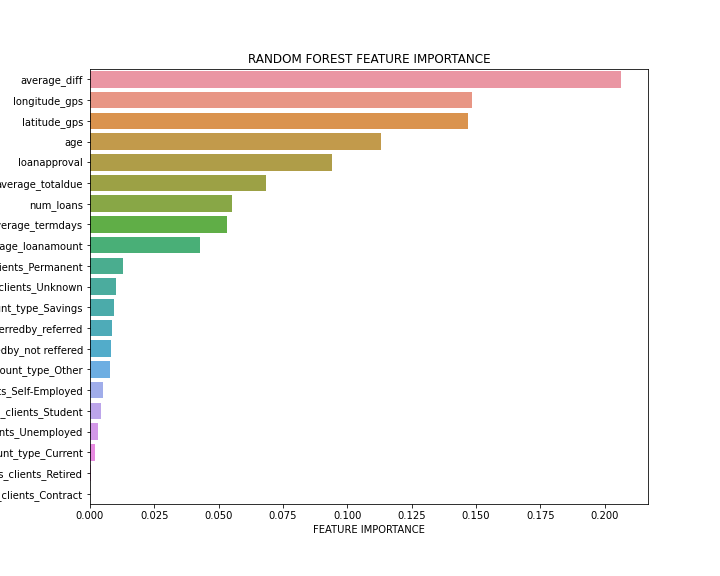
* Scaling of numerical features
* Encoding of the categorical features
* Upsampling of the minority class - The dataset was highly imbalanced in term of the binary class target variable with good loans being 78% of the data and bad loans being only 22%
* Splitting data into the train and test set - the train set was 70% while the test set was 30%

Models tested include a support vector classifier, a random forest classifier as well as an xgboost classifier. Models were evaluated based on their accuracy score as well as their roc\_auc\_score. The best performing model was the random forest classifier with an accuracy score of 93% and an roc\_auc\_score of 0.977.

The resulting confusion matrix of the model was as follows:

|  |  |  |
| --- | --- | --- |
|  | Predicted Negative | Predicted Positive |
| Actual Negative | 1005 | 36 |
| Actual Positive | 106 | 903 |

The random classifier ranked the feature importances as follows:



The most important features as identified by our model include the following: average diff (which is the average time taken by a customer in repaying their loans), location data, age, number of loans taken, employment status, average total due as well as the time taken to get the loan application approved.

**RECOMMENDATIONS**

With the identified features, customers can be profiled based on the following characteristics:

1. Time taken to make the first payment of the loan. By how many days are customers early or late based on the first due date?
2. Location data: Cluster customers based on their location and monitor if indeed customers from specific regions are more likely to default on their loans. If true, mark customers from this “hot-zones”.
3. Age : Cluster customers based on ages and the default rate of each cluster.
4. Number of loans borrowed by the customer.
5. Average time taken for past loans to get approved. Further investigation on why certain loans take longer to get approved than others.
6. Employment status: Customers with a permanent job have a better repayment history.
7. Loan amounts borrowed: Customers who routinely take loans of smaller amounts have a better repayment history.

For future analysis and more accuracy in the data, more data could be provided on the clients. As the model had so many unknown features.

In terms of the prediction model, the model can be optimized using hyperparameter tuning for better accuracy. Tests on whether dimension reduction can improve model performance should also be performed. As the data is heavily imbalanced, other methods of dealing with imbalanced classification can be tested such as undersampling the majority class as well as SMOTE.