School of Informatics & IT
TEMASEK POLYTECHNIC

Big Data Programming
CBG1C04
Report

Submitted by

LIM YUAN HER (1780113E)

05 March 2018

Declaration of Work Originality Template

Specialist Diploma in Big Data Management
Big Data Programming (CBG1C04)
AY2017/2018 Oct Semester
Assignment

Intake: April 2017

Submitted by: 1780113E LIM YUAN HER

Date: 05/03/2018

"By submitting this work, I am declaring that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person's work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence and that disciplinary action will be taken for plagiarism."


Lim Yuan Her

Name and Signature of student: ……..

Table of Contents

List of Tables

Table of Figures

1. Introduction

The report summarizes the design and implementation details of data preparation/ exploration (using Spark SQL) and model building (using Spark ML) on the Lending Club loans data set. This report is divided into the following sections:

Section 2: Describes the application domain i.e. Lending Club.

Section 3: Describes the research goals.

Section 4: Describes the data source/ schema information including the source where the dataset is downloaded from.

Section 5: Describes the data cleaning/ transformation/ integration steps performed on the dataset and the results of the data exploration performed.

Section 6: Describes the model built and rationale for choosing that model. The evaluation results of the model are also detailed.

Section 7: Describes the insights gained from performing the analytics.

Section 8: Describes how the solution is deployed in a production environment and any commercial benefits from adopting this solution.

Section 9: Describes any improvements/ suggestions for further work.

Section 10: Lists any external reference material used.

2. Application domain

2.1 Introduction

Lending Club is an online crowdfunding platform for peer to peer lending, facilitating personal loans, business loans, and financing.

Borrowers access loans through an online or mobile interface and investors provide capital in exchange for earning interest (peer-to-peer (P2P) lending).

Being an online-only operation results in cheaper operating costs and overheads, thus this offers lenders higher returns compared to traditional bank products. Borrowers can borrow money at lower interest rates, even after accounting for platform and credit checking fees.

Interest rates are set by lenders who compete for the lowest rate based on a reverse auction model or a fixed rate based on borrower's credit profile.

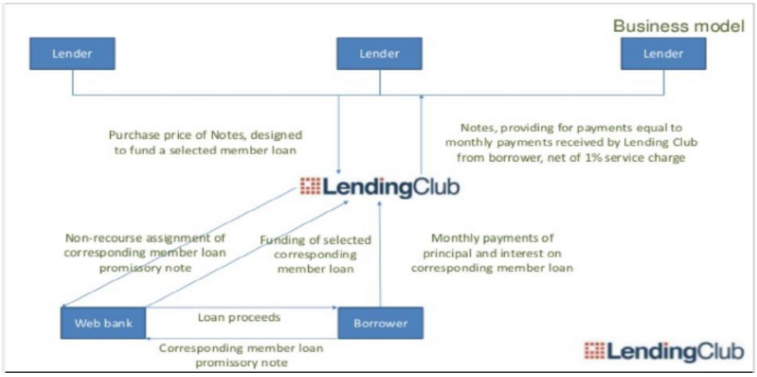


Figure 1 – Lending Club's Business Model

3. Research Goals

The overriding concern on the LendingClub crowdfunding platform is the risk of defaults on the loan(s) being offered, thus resulting in financial losses for the lenders. Thus, any solution that is available to mitigate this risk would be greatly beneficial to the lenders, who could be notified in advance whether a particular loan is susceptible to default.

The research goal of this report is to attempt to predict the risk of the loan being default based on past loans data using partial loans data from 2017 (Quarters 1 to 3).

Logistic Regression/ Decision Tree classification algorithms are proposed to be evaluated with the goal of maximizing AUC (areaUnderROC) and to evaluate the prediction performance.

4. Data Source/Schema

4.1 Data Source

The Dataset is derived from 2 input files:

- "LoanStats.csv" - loans data for all loans issued including current loan status and payment information
- "LCDataDictionary.xlsx" - Data Dictionary containing definitions for all data attributes

Each loan record (15,000 records in loan data file) has 145 attributes (columns) and Data Dictionary has 2 attributes (attribute name, description)

4.2 Data Schema

There are 7 loan statuses in the "loan_status" attribute: Charged Off, Current, Default, Fully Paid, In Grace Period, Late (16-30 days), Late (31-120 days). For the purpose of model building and validation, the "Charged Off" (default loan) and "Fully Paid" (non-default) statuses are used.

5. Data Preparation

5.1 Data Cleaning

The first step in data analytics is to prepare the data. This involves the following:

1. Drop columns with >50% missing values
    - Columns with more than half of null/missing values will not provide any meaningful insights as these usually occur if they are dependent on the values entered in other columns or they are optional fields that are not mandatory for the user to enter. As such, such columns are removed from our dataset prior to further analysis.
2. Drop columns with strong correlations with each other
    - Columns that have strong correlations with each other are normally information that is related and for data analysis purpose, only one of them would be sufficient. Columns with 90% correlation are removed from the dataset.
3. Remove columns of date datatype
    - As our data analysis involves binary classification i.e. prediction of loan defaults, time series i.e. date type columns are not relevant and

as such, are removed from the dataset. This reduces the number of column data needed to be processed.

4. Remove loan records with statuses other than "Fully Paid" or "Charged Off"
    - As we are only interested in fully paid (“Fully Paid”) and defaulted (“Charged Off”) loans, other records with loan statuses other than the aforementioned (in “loan_status” column) are removed.
5. Remove duplicate records
    - To prevent misleading analysis results as a result of record duplication due to erroneous entries, such records are removed from the dataset.
6. Remove records having missing values in any of the columns
    - As our data analysis requires selection of features to be used for Logistic Regression model building, we require complete entries in each column. Thus, records with missing entries in any of the columns will be removed from the dataset.
7. Check Outlier Data
    - Based on the outlier data analysis, the number of outlier records as a percentage of total records is not significant for most columns except for "tot_coll_amt" column, which, when visualized using a histogram, shows that the records distribution is skewed towards zero values occurring in majority of the records. No further action is taken.
8. Check for distinct values in columns
    - When checking the number of distinct values in each column, the following were observed:
        i. The "emp_title" column have too many inconsistent entry values e.g. multiple title description for potential same job e.g. "Accounts Payable"/ "Accounts Payable Clerk", inconsistent capitalization for titles with multiple wording e.g. "Accounts Receivable"/ "Accounts receivable", ambigious titles e.g. "Admissions", "Assembly", "BMB". As it is not useful, this column is dropped from the dataset.

ii. The "emp_title" and "zip_code" columns have >1000+ distinct values, and thus, is removed as they will not provide any meaningful insight.

iii. The "pymnt_plan", "hardship_flag", "debt_settlement_flag" columns each have only 1 distinct value. Thus, they are dropped due to the limited useful insight they can provide.

The table below summarizes the number of records and columns pre- and post- data cleaning:

Description	Before	After
Number of Records	324937	13555
Number of Columns	145	80

Table 1 – Pre- and Post- Data Cleaning Dataset Statistics

Note that only loan records with loan status of “Fully Paid” and “Charged Off” are included in the cleaned dataset.

5.2 Data Transformation/ Integration

After the data cleaning process, before the dataset is input to the learning model, the loan default predictor variable i.e. “loan_status” column has to be encoded using a numerical representation. This is necessary for the classification algorithm to classify the predictor variable i.e. Loan Status for evaluation:

s/n	Loan Status	Numerical Representation
1.	Fully Paid	0
2.	Charged Off	1

Table 2 – Loan Status Encoding

5.3 Data Exploration

5.3.1 Descriptive Statistics

The first step in exploratory data analysis is to extract basic statistical information about the various features (columns) in the dataset. The “describe” command is used to achieve this on both numerical and categorical columns.

5.3.1.1 *Numerical Columns*

	loan_amnt	int_rate	annual_inc	dti	delinq_2yrs	inq_last_6mths	mths_since_last_delinq	open_acc
count	13555.000000	13555.000000	1.355500e+04	13555.000000	13555.000000	13555.000000	13555.000000	13555.000000
mean	14156.029141	0.149561	8.779431e+04	18.610225	0.613722	0.668462	34.511177	12.884618
std	9398.831175	0.058947	9.265269e+04	8.606561	1.175517	0.898249	22.023618	5.855155
min	1000.000000	0.053200	9.000000e+03	0.000000	0.000000	0.000000	0.000000	2.000000
25%	7000.000000	0.113900	5.500000e+04	12.620000	0.000000	0.000000	16.000000	9.000000
50%	12000.000000	0.139900	7.500000e+04	17.830000	0.000000	0.000000	31.000000	12.000000
75%	20000.000000	0.179900	1.030000e+05	23.975000	1.000000	1.000000	50.000000	16.000000
max	40000.000000	0.309900	8.300000e+06	195.240000	19.000000	5.000000	160.000000	72.000000

pub_rec	revol_bal	...	num_rev_accts	num_tl_120dpd_2m	num_tl_30dpd	num_tl_90g_dpd_24m	num_tl_op_past_12m
13555.000000	13555.000000	...	13555.000000	13555.000000	13555.000000	13555.000000	13555.000000
0.245592	14984.175876	...	15.320472	0.001475	0.007304	0.152785	2.689709
0.638483	21291.762820	...	8.366229	0.038385	0.089379	0.698160	2.061310
0.000000	0.000000	...	2.000000	0.000000	0.000000	0.000000	0.000000
0.000000	5408.000000	...	9.000000	0.000000	0.000000	0.000000	1.000000
0.000000	10221.000000	...	14.000000	0.000000	0.000000	0.000000	2.000000
0.000000	18092.000000	...	20.000000	0.000000	0.000000	0.000000	4.000000
11.000000	651124.000000	...	101.000000	1.000000	3.000000	19.000000	20.000000

Figure 2- Numerical Columns Descriptive Statistics

From the results above, the following can be observed:

i. The average loan amount is $14.2K and loans are offered at an average interest rate of 15% with the maximum topping at 30.1%.

ii. The average annual income of a borrower is approximately $88K with minimum of $9K and maximum at $8.3M.

5.3.1.2 *Categorical Columns*

	term	grade	sub_grade	emp_length	home_ownership	verification_status	purpose	title	addr_state	initial_list
count	13555	13555	13555	13555	13555	13555	13555	13555	13555	
unique	2	7	35	12	5	3	12	12	49	
top	36 months	C	C5	10+ years	MORTGAGE	Source Verified	debt_consolidation	Debt consolidation	CA	
freq	10319	4733	1092	5282	7576	5962	7559	7560	1878	

initial_list_status	application_type	disbursement_method	loan_status
13555	13555	13555	13555
2	2	2	2
w	Individual	Cash	Fully Paid
9415	12969	13549	11993

Figure 3- Categorical Columns Descriptive Statistics

From the results above, it is observed that the majority of the loans that are subsequently fully paid or defaulted are individual 36-month loans using the cash disbursement method. There is a much higher proportion of loans that are fully paid (88.4%) as compared to defaulted loans and employees with

more than 10 years of service carrying home mortgages tend to make up slightly higher proportion of borrowers with debt consolidation as their stated purpose.

5.3.2 Histograms (distribution profile of numerical features)

Next, the skewness of each numerical column is checked using histogram plots and calculating the skewness coefficients.

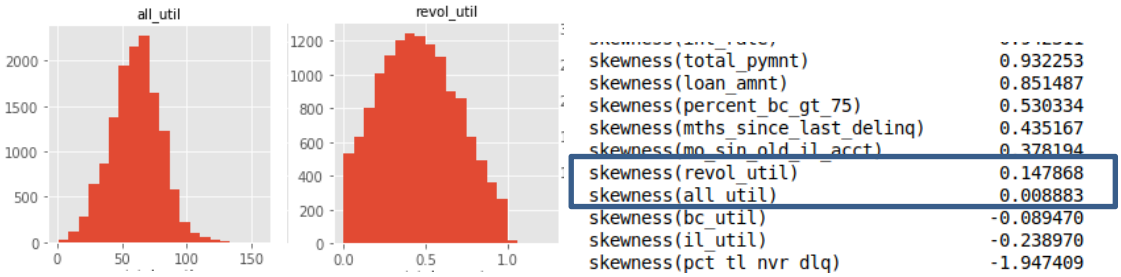


Figure 4- Skewness Check

From the histogram plots and calculation of the skewness coefficient for each numerical column, it is observed that except for "all_util" and "revo_util" columns which exhibits gaussian distribution characteristics (see above figures), all others exhibit lognormal characteristics and depending on the learning model used for prediction, data normalization may be required.

5.3.3 Correlation Analysis

Next, the correlation between numerical columns is analyzed and the following are observed:

i. The following table summarizes columns that have particularly strong correlation as compared with other columns:

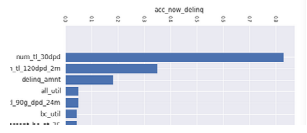
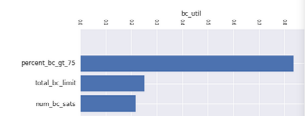
s/n	Column	Strongly correlated with	
1.	acc_now_delinq	num_ti_30dpd	
2.	bc_util	percent_bt_gt_75	

Table 3 – Listing of strongly correlated columns

i. “tax_liens” feature has correlation only with the “total_bc_limit” feature.

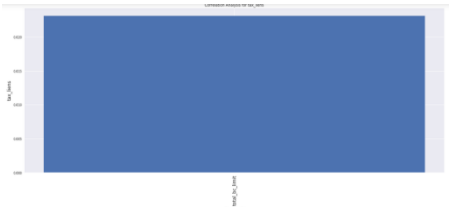


Figure 5- “tax_liens” correlation

ii. "total_bc_limit" is the only feature that exhibits some correlation with all other numerical features.

iii. "policy_code", "out_prncp" and "loan_amnt" features do not exhibit correlation with all other numerical features.

5.3.4 Box/ Density Plots (correlation of loan status with numerical features)

Next, the correlation between the numerical columns with the “loan_status” column (for "Charged Off" and "Fully Paid" instances) is analyzed using box and density plots and the following are observed:

1. Occurrences of defaulted loans tend to be associated with higher interest rates offered whilst that for fully paid loans is associated with lower interest rates.

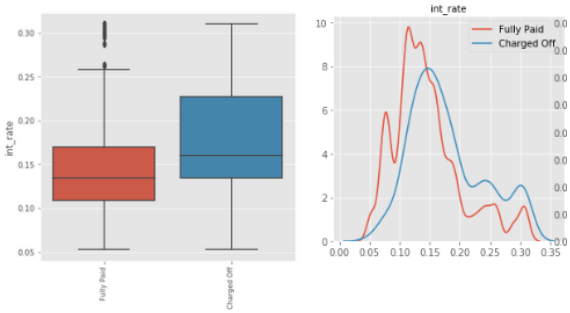


Figure 6 – Interest Rate-Loan Status Correlation

2. Occurrences of defaulted loans tend to be associated with much lower total payment amounts recorded (“total_pymnt”) whilst that for fully paid loans tend to be associated with much higher total payment amounts recorded.

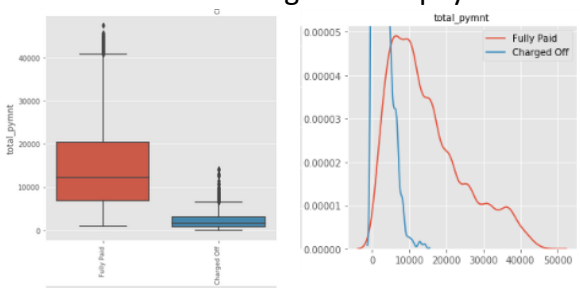


Figure 7 – Total Payments-Loan Status Correlation

5.3.5 'Mosaic Plots (correlation between categorical features)

Next, the correlation between categorical columns is checked using mosaic plots.

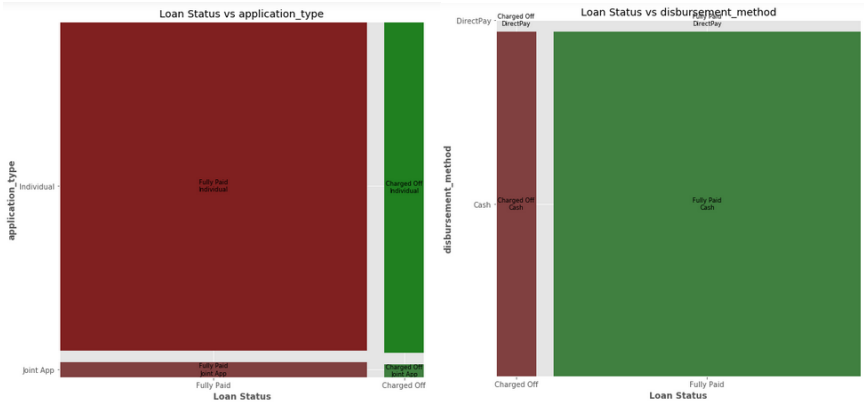


Figure 8 – "application_Type", "disbursement_method" correlation with "loan_status"

It is observed that the horizontal splits occur at similar locations on the y-axis indicating no strong relationship exists between loan status and "application_type", "disbursement_method" features. For the other features, the horizontal splits do not occur at different levels significant enough to warrant a strong correlation between these and the "loan_status" feature.

5.3.6 Heat Maps (correlation of categorical feature groups with numerical features)

Next, we check the correlation for each categorical feature group (with "loan_status" column) against the median values of the respective numerical features with the following observations:

1. If the loan grade ("grade") is of lower quality, the interest rates tend to be higher for both "Fully Paid" and "Charged Off" loans.

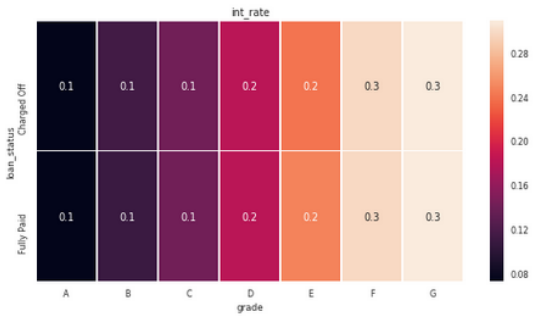


Figure 9 – Loan Grade-Interest Rate correlation with Loan Status

2. Total payments ("total_pymnt") tend to be lower for "Charged Off" (defaulted) loans as compared to "Fully Paid" loans for all types of loan grades ("grade"), employment tenure ("emp_length"), home ownership status ("home_ownership"), loan verification status ("verification_status"), initial list status ("initial_list_status"), application type ("application_type") and disbursement method ("disbursement_method").

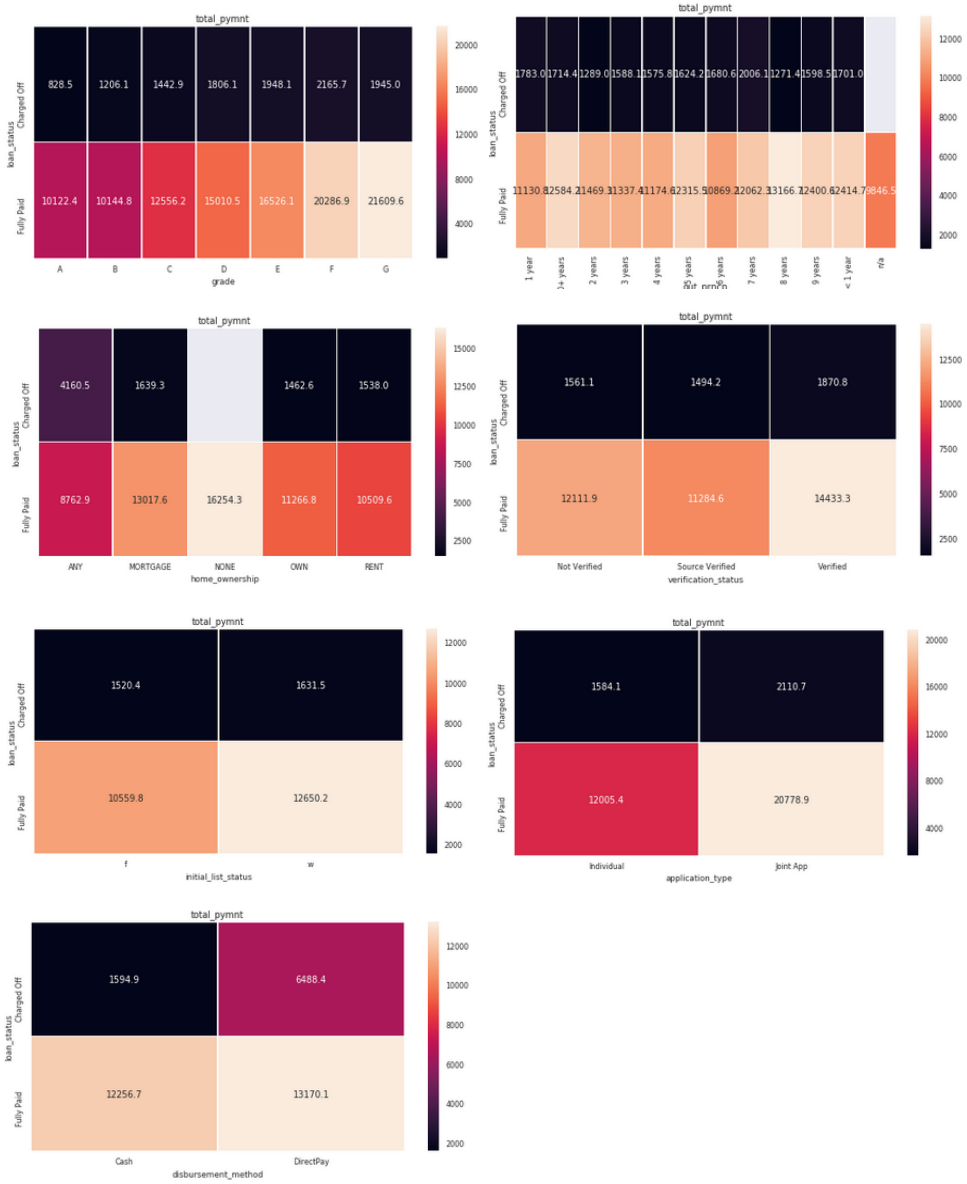


Figure 10 – Total Payments correlation with other features

3. For all categories of "emp_length", interest rates are higher for "Charged Off" loans.

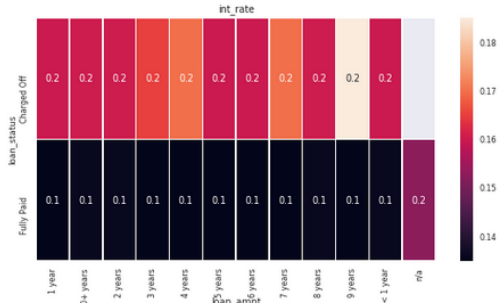


Figure 11 – Employment Tenure-Interest Rate correlation with Loan Status

4. Loan amounts ("loan_amnt") tend to be higher for "Charged Off" loans using "DirectPay" disbursement method and of "ANY" home ownership status.

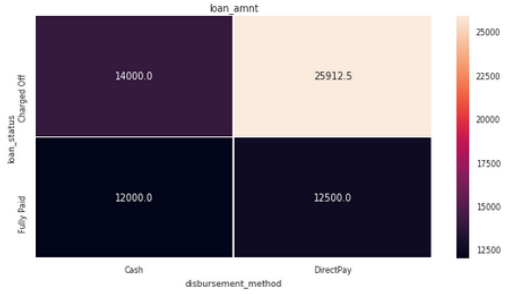


Figure 12 – Loan Amount-Disbursement Method correlation with Loan Status

5. "out_prncp" tends to have no correlation with any of the categorical features.

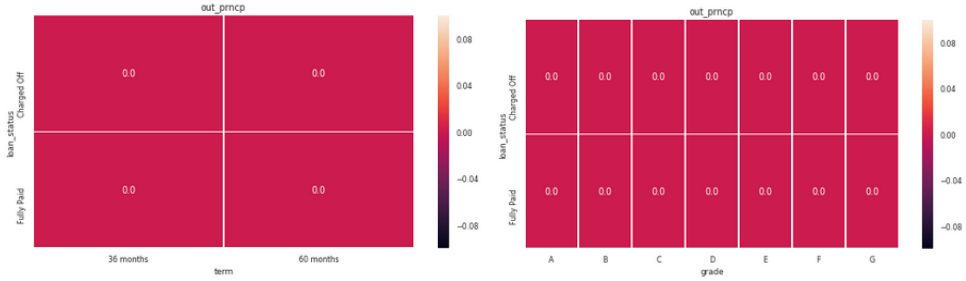


Figure 13 – "out_prncp" correlation with other features

5.3.7 Pareto plots (distribution profile for categorical features)

Next, the occurrence frequency for each categorical feature is analyzed using pareto charts (based on the Pareto Principle (80/20 Rule) to uncover the 20% of causes that are creating 80% of the problems) and pie charts. The following observations were made:

1. The majority of loans (88.5%) are fully paid loans type and approximately 1/10th of the total loans defaulted.

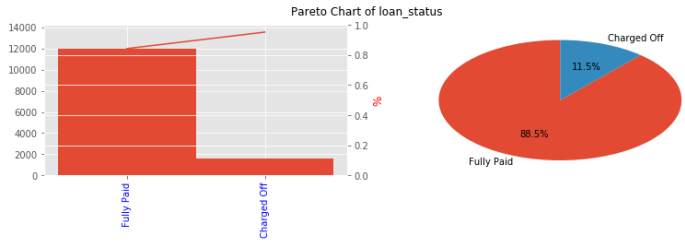


Figure 14 – Loan Status Distribution Profile

2. Approximately 3/4 of the loans are of shorter tenure (36-month) while only 1/4 are of longer tenure (60-months).

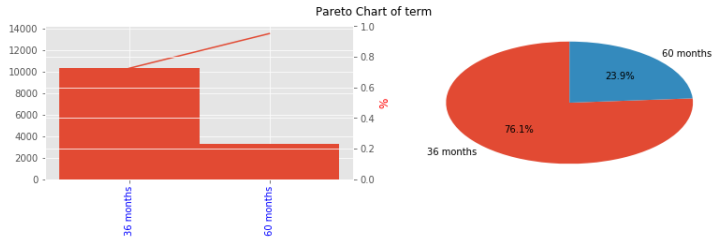


Figure 15 – Employment Tenure Distribution Profile

3. The majority of borrowers are either on home mortgages or rental.

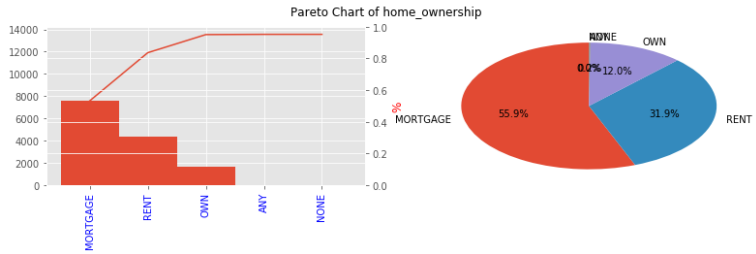


Figure 16 – Home Ownership Type Distribution Profile

4. "Debt Consolidation", "Credit card refinancing" and "Home improvement" are the top 3 reasons given for the loans.

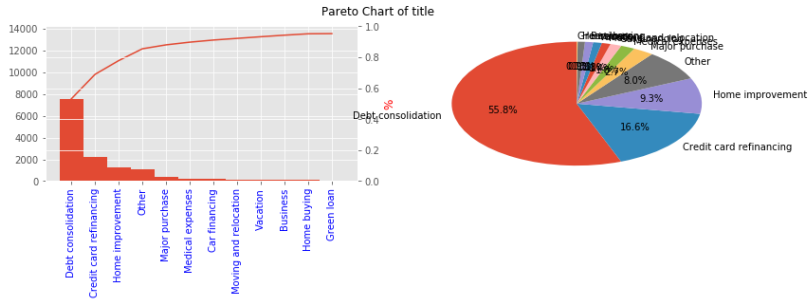


Figure 17 – Loan Purpose Distribution Profile

5. The majority of the loans are individual application type.

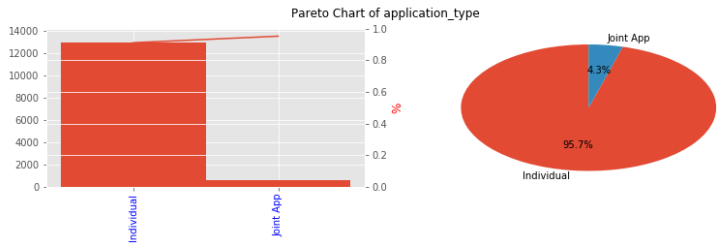


Figure 18 Individual/Joint App Distribution Profile

6. Model Building

6.1 Classification Algorithm Selection

For the purpose predicting loan default from the loans data, there are 2 classification algorithms that are suitable for use, namely Decision Tree and Logistic Regression. The idea is to fit the logistics regression or Decision Tree model using binary classification to train and predict the Loan Status (Fully Paid (0.0) and Charged Off (1.0)) for the dataset.

In this case, the performance of these 2 algorithms is compared with and without tuning using k-fold cross-validation in combination with grid search to find the optimal combination of the tuning parameters i.e. regularization parameter in Logistic Regression or the depth parameter for a Decision Tree. The number of folds used in cross-validation for model tuning is also varied to investigate the performance improvements (if any) of using a higher value.

6.2 Feature Selection

Feature selection is a process of automatically selecting those features that contribute most to the prediction variable or output of interest. The benefits of performing feature selection before modelling are:

1. Reduces Overfitting: Less redundant data means less opportunity to make decisions based on noise.
2. Improves Accuracy: Less misleading data means modelling accuracy improves.
3. Reduces Training Time: Less data means that algorithms train faster.

6.2.1 Determination of number of features to include

To determine how many features to be used in the model training, the dimensionality reduction aspect of Principal Component Analysis (PCA) is used with the following results:

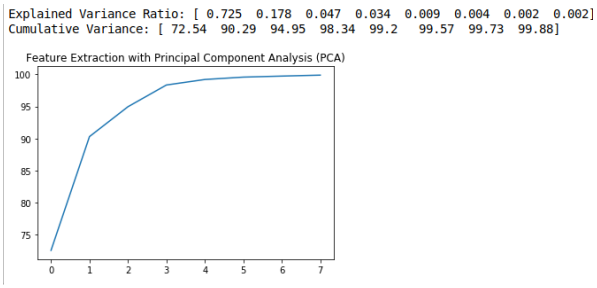


Figure 19- PCA Variances

Based on Figure 2 above, it is observed that 4 principal components are sufficient to explain 98% of the full variance i.e. 4 features are sufficient to input to the training model.

6.2.2 Determination of features to use

Based on the exploratory data analysis conducted in the previous sections, it is proposed to use the following features for the machine learning model:

"int_rate", "loan_amnt", "total_pymnt", "out_prncp", "policy_code"

“int_rate” and “total_pymnt” features are chosen based on the correlation analysis of loan status with numerical features in section 5.3.4 and "loan_amnt", "out_prncp" and "policy_code" features are chosen due to their weak correlation with other features.

6.3 Model training and performance evaluation

The dataset is split into training data (70%) and test data (30%) randomly. In this case, the random split generated 4001 test records (3553 fully paid, 448 defaulted) and 9554 training records (8440 fully paid, 1114 defaulted).

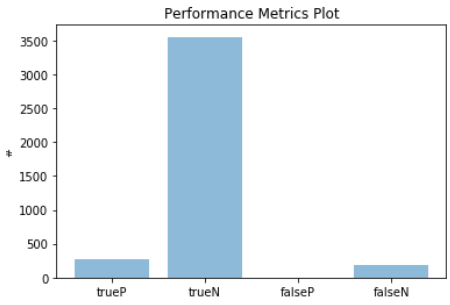
The table below summarizes the performance evaluation using Logistic Regression and Decision Tree (with and without k-fold cross-validation):

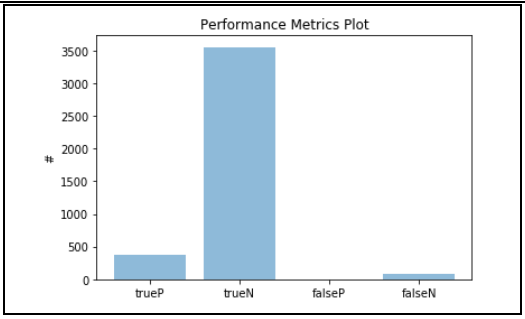
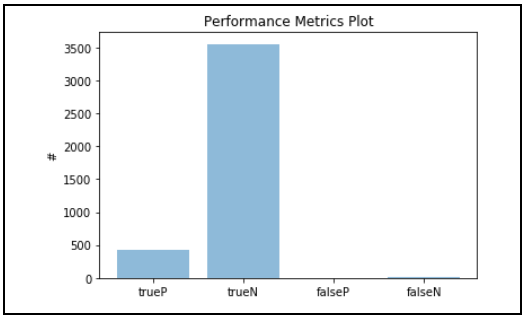
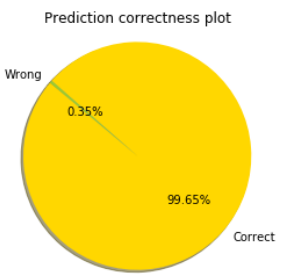
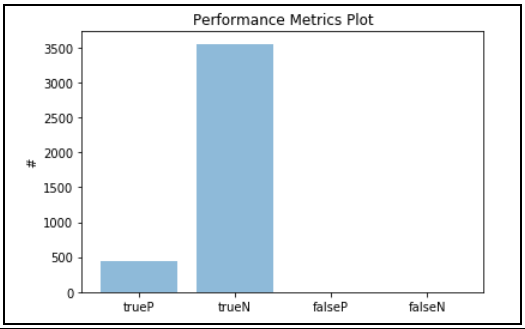
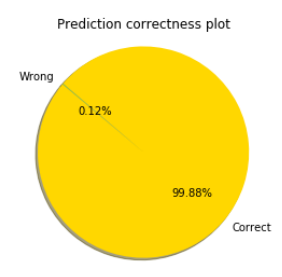
Algorithm	AUC	AUPR	Accuracy	Error	Optimal Grid Search Hyperparameters			
					Logistic Regression			Decision Tree
					regParam	elasticNetParam	maxIter	maxDepth
Logistic Regression	0.9969	0.9884	95.53%	4.47%	0.01	0.01	50	-
Logistic Regression (k-fold cross-validation) n =3	0.9971	0.9948	98.03%	1.97%	0.01	1	50	-
Logistic Regression (k-fold cross-validation) n =10	0.9971	0.9948	98.03%	1.97%	0.01	1	50	-
Decision Tree	0.9951	0.9895	99.65%	0.35%	-	-	-	5
Decision Tree (k-fold cross-validation) n = 3	0.9972	0.9943	99.88%	0.12%	-	-	-	6
Decision Tree (k-fold cross-validation) n = 10	0.9972	0.9943	99.88%	0.12%	-	-	-	6

Table 4 - Classification Algorithm Performances

1. From table 2 above, it is observed that using k-fold cross-validation with grid search for hyperparameter tuning improves the performance for both Logistic Regression and Decision Tree algorithms, but increasing the number of folds does not improve the performance any further.
2. It is observed that using Decision tree algorithm outperforms that using Logistic Regression (with or without using k-fold cross-validation).

The pie/ bar charts below summarize the prediction metrics (True Positive/ True Negative/ False Positive/ False Negative) for each algorithm performance:

Algorithm	Performance Metrics Plot	Correctness Plot
Logistic Regression		

Logistic Regression (3-fold cross-validation) Logistic Regression (10-fold cross-validation)		
Decision Tree		
Decision Tree (3-fold cross-validation) Decision Tree (10-fold cross-validation)		

Note: Table

trueP True positives are how often the model correctly predicted a loan defaulted

falseP False positives are how often the model predicted a loan defaulted when it was fully paid

trueN True negatives indicate how the model correctly predicted a loan was fully paid

falseN False negatives indicate how often the model predicted a loan was fully paid when in fact it was defaulted

Table 5 - Classification Algorithm Metric/ Correctness Plots

From Table 3 above, it is observed that the false negatives occurrences is much lower when using Decision Tree thus improving the overall performance as compared to that for Logistic Regression.

6.4 Model back and forward testing

Based on the model using Decision Tree algorithm (maxDepth = 6) evaluated as providing the best performance, back and forward testing of the model performance is evaluated using loans data available from periods prior and post to that used for the model building and evaluation, details are as follows:

Data used for model building : 2017Q1 to 2017Q3

BackTesting Data : 2016Q1, 2016Q3, 2016Q4

Forward Testing Data : 2017Q4

The performance is evaluated by checking the quantity of True and False Positives/Negatives as a proportion of actual total Positives/Negatives. Refer to table 5 above for the definitions of True and False Positives/Negatives.

Back/Forward Testing Data Period	Total Positives	Total Negatives	True Positives	False Positives	True Negatives	False Negatives
2016Q1	10062	47469	10062 (100%)	0 (0%)	41833 (88.13%)	5636 (11.87%)
2016Q3	6869	27345	6869 (100%)	0 (0%)	25538 (93.39%)	1807 (6.61%)
2016Q4	4823	21511	4823 (100%)	0 (0%)	21205 (98.58%)	306 (1.42%)
2017Q4	25	3846	25 (100%)	0 (0%)	3843 (99.92%)	3 (0.08%)

Table 6 – Model Back/Forward Testing Performance Results

From table 6 above, it can be observed that the model used is able to predict all the defaulted loans correctly (100% True Positives). However, the model is susceptible to predict loans as fully paid i.e. not defaulted, when in fact, the loans were defaulted (False Negatives are not zero).

In addition, the model performance improved substantially nearer to the period used for model building/evaluation (2017Q1 to 217Q3). For example, the number of False Negatives decreased from 2016Q1 to 2016Q4 and was very low for 2017Q4. This suggests that the model needs to be re-evaluated periodically e.g. every 3-6 months to ensure relevancy and ability to capture changes in data pattern.

7. Analytic Insights

7.1 Summary

The following summarizes the insights derived from the exploratory data analysis and model building performed in the previous sections:

i. The average loan amount offered is $14.2K and loans are offered at an average interest rate of 15% with the maximum at approximately 30%.
ii. The average annual income of a borrower is approximately $88K with lowest of $9K and highest at $8.3M.
iii. Majority of the loans are individual 36-month loans using cash disbursement method with a higher proportion being fully paid (88.4%) as compared to defaulted loans. 0
iv. Employees with more than 10 years of service with home mortgages make up a slightly higher proportion of borrowers with debt consolidation indicated as the stated purpose for the loans.
v. Defaulted loans tend to be associated with higher interest rates offered and with much lower total payments received to date for total amount funded.
vi. Interest rates offered are higher for lower grade loans.
vii. Total payments tend to be lower for defaulted loans for all loan grades, employment tenure, home ownership status, loan verification status, initial list status, application type and disbursement method.
viii. Interest rates offered are higher for defaulted loans regardless of employment tenure.
ix. Loan amounts are higher for defaulted loans using the "DirectPay" disbursement method and non-stated home ownership status.
x. Approximately 3/4 of the loans are of shorter tenure (36-month) while only 1/4 is of longer tenure (60-months).
xi. The majority of borrowers are either on home mortgages or rental.
xii. “Debt Consolidation”, “Credit card refinancing” and “Home improvement” are the top 3 reasons indicated for the loans.
xiii. Using Decision Tree algorithm to build a machine learning model for loan status default prediction outperforms that using Logistic Regression (regardless of whether k-fold cross-validation with grid search is used or not).

8. Deployment

After the model has been built and its performance validated, the Pipeline definition can be saved for later use, including the pipeline structure and all the definitions of all the Transformers and Estimators.

The following diagram illustrates the process where, in a production environment, how the solution proposed can be deployed:

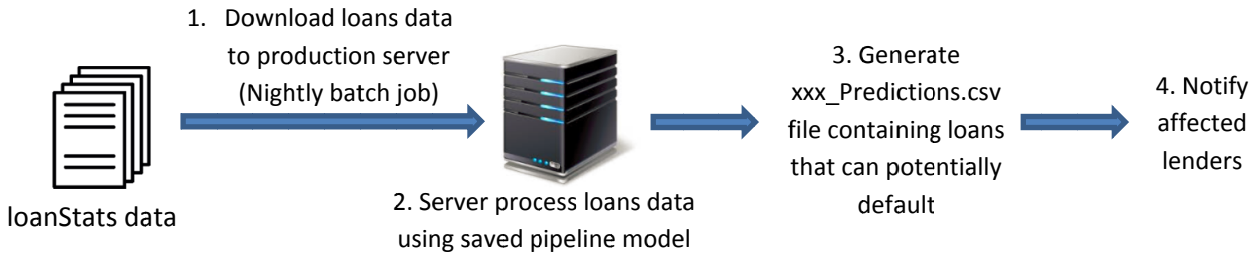


Figure 20 – Proposed Deployment Workflow

The solution involves downloading the loanStats data to the production server nightly, and the server processes the loans data using the saved pipeline model and generates a xxx_Predictions.csv file (where xxx refers to the original loans data filename) highlighting loans that could potentially default. This information can then be further processed for dissemination to the affected lenders.

9. Conclusion

In this report, an analytics solution for LendingClub has been developed with the use of machine learning techniques (Logistic Regression/ Decision Tree) to extract useful insights on the loans data by performing exploratory data analysis and to predict the risk of loan defaults. Further improvements that could be implemented in future include:

1. Migrating this server-based solution to a cloud-based one
2. integrating with real-time loans data from LendingClub to perform the analysis and prediction instead of nightly batch download of loans data in csv format

10. References

1. Lending Club Loans Data (https://www.lendingclub.com/info/download-data.action)