Sam Loyd

SBA Loan – Final Report

November 2020

Executive Summary

The purpose of this project was to analyze the Small Business Administration data set containing information on small business loans from 1961 to 2014 in order to provide predictions on two key features. Predicting loan repayment was the primary objective since these loans are backed by the U. S. government. A secondary consideration for predicting job creation was included. A successful model would allow for the selection of repaid loans providing the highest returns in the form of new jobs.

The analysis concluded that most loans were repaid. It also discovered that most of the loans had created no jobs. There were several examples of high value outliers in most features that made the analysis more challenging. Methods to minimize the impact of those values were utilized. Features were excluded to minimize concerns of racial bias. After the data was understood, it was cleansed and prepared for modeling. Ultimately, two models were selected. One was selected to create a prediction for loan repayment and the other for the number of jobs a loan would create. Each of these were trained using the same subset of data. Testing and validation of their ability to predict those values was performed against the same two remaining samples of data.

The ability to predict loan repayment proved very successful. It was also the most straight forward of the two models to work with. The job creation model, while successful at explaining most variance in predictions left some concerns given the seasonal or time-based data that was required to achieve positive results. This model will require additional monitoring and may prove troublesome in a production environment due to drift caused by changing data. More research is required against current data to validate its success.

Abstract

The loans in this data set were provided by the American Statistical Association (ASA) as a teaching tool in introductory statistics and uploaded to Kaggle (2020). This data once understood and prepared in a usable format was used to train predictive models for use against future data sets. The primary goal of this project was to determine the likelihood of loan repayment. A secondary goal was to provide a value predicting the number of new jobs each loan request is likely to create.

Introduction - Background of the Problem

The funds used to provide small business loans are commonly backed by the government through the Small Business Administration (SBA). Up to eighty-five percent of a default can be repaid by the Small Business Administration to the banking institution holding the loan (SBA, 2011). In that spirit, predictions were desired to make sound selections. The purpose of this analysis was to look at two target variables to not only determine loan repayment, but to also maximize the number of new positions created per dollar invested. Predictions could then be used to target future loans to businesses that are likely to repay their debt while at the same time providing a recommendation allowing for the largest return to the taxpayer in the form of new jobs.

Before data preparation the data set consisted of twenty-seven variables. There were ten features that focused on the borrower and included descriptive information about the business. These were *loanNr_ChkDgt, name, city, state, zip, noemp, newexist, urbanrural, NAICS* and *franchisecode*. *LoanNr_ChkDgt* serves as a unique identifier and as such was not part of the analysis. *Noemp* provided a current count of employees at the time of the loan request.

Newexist indicated whether this was an existing business, and *urbanrural* indicated the setting. *NAICS* and *franchisecode* provided industry standard coding and franchise coding information. The next grouping of features focused on the lending institution. *Bank* and *bankstate* fell into this category. The remaining fifteen features provided data on individual loans. Two of these, *MIS_status* and *createjob,* were the target variables. *Chgoffpringr* and *chgoffdate* provided information about a default and had to be guarded against providing future data to any model (Kagan & Brack, 2020). *Balancegross* carried similar concerns. *SBA_appv* indicated the amount of risk carried by the Small Business Administration. The entire loan amount requested of the banking institution was provided in *grappv. Term* indicated the repayment time allocated for the loan given in months. The date and fiscal year the loan was approved was provided in *approvaldate* and *approvalfy. RevLineCr* indicated whether the loan was part of a revolving line of credit. If the loan was part of a low documentation program, that was indicated in the *lowdoc* feature. Disbursement amounts and dates were noted in the *disbursementamount* and *disbursementdate* features. And finally, the number of jobs retained over the loan was also indicated by the *retainedjob* feature. This also carried concerns of future information. Data was included for 899,164 total loans.

The following chart shown in figure one was taken from a supplemental attachment included with the data set that shows original data types and a brief description.

Figure 1

SBA Descriptions

Table 1(a). Description of 27 variables in both datasets.

Variable name	Data type	Description of variable
LoanNr_ChkDgt	Text	Identifier – Primary key
Name	Text	Borrower name
City	Text	Borrower city
State	Text	Borrower state
Zip	Text	Borrower zip code
Bank	Text	Bank name
BankState	Text	Bank state
NAICS	Text	North American industry classification system code
ApprovalDate	Date/Time	Date SBA commitment issued
ApprovalFY	Text	Fiscal year of commitment
Term	Number	Loan term in months
NoEmp	Number	Number of business employees
NewExist	Text	1 = Existing business, 2 = New business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise code, (00000 or 00001) = No franchise
UrbanRural	Text	1 = Urban, 2 = rural, 0 = undefined
RevLineCr	Text	Revolving line of credit: Y = Yes, N = No
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement date
DisbursementGross	Currency	Amount disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan status charged off = CHGOFF, Paid in full = PIF
ChgOffPrinGr	Currency	Charged-off amount
GrAppv	Currency	Gross amount of loan approved by bank
SBA_Appv	Currency	SBA's guaranteed amount of approved loan

Note. From "Should This Loan be Approved or Denied? A Large Dataset with Class Assignment Guidelines" by Li, M., Mickel, A. & Taylor, S., 2018, Journal of Statistics Education, 26(1), p. 56.

Methods

The CRISP-DM methodology was implemented for this project (Siegel, 2016). As such, six iterative stages were followed.

- Business Understanding

- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

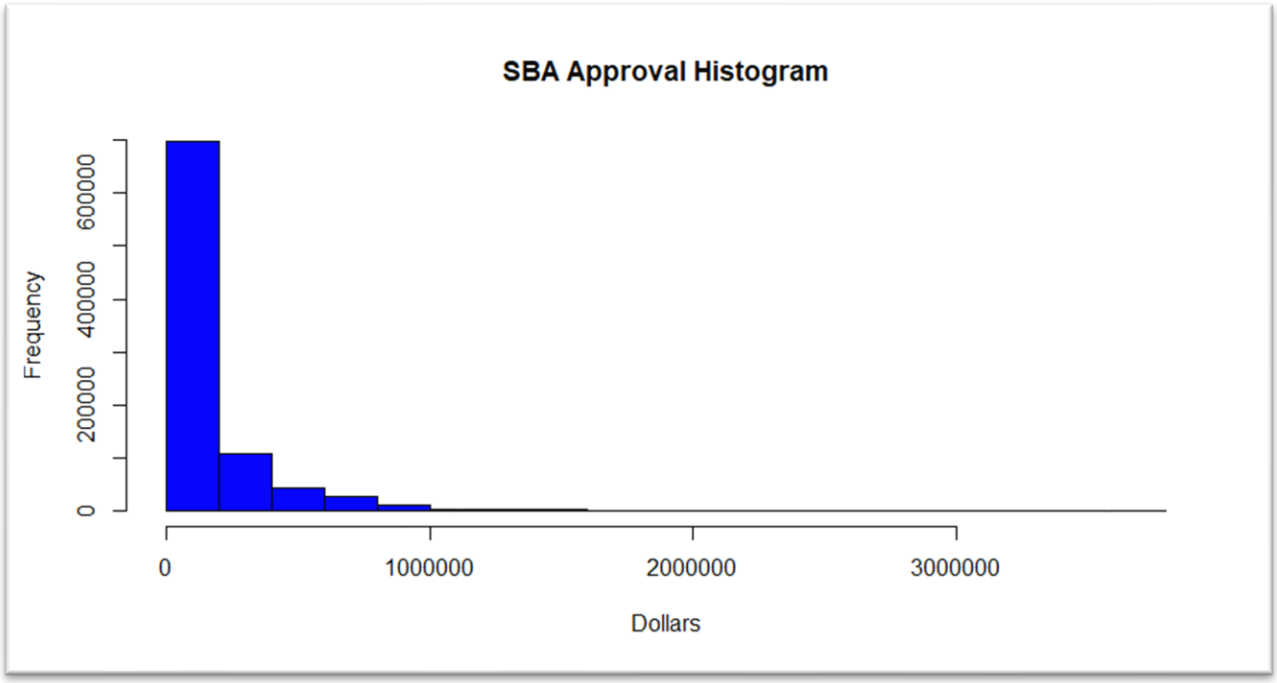
In the context of this paper, the first four will be covered in the current methods section. The evaluation and deployment phases of CRISP-DM can be found in the results and conclusion sections.

In the business understanding phase, domain research was conducted. Here there were several findings that influenced the data and progression of this project. The first involved a consideration of racial bias. Using local geographic information such as a zip code for loan approval brings significant moral and legal concerns (Siegel, 2016). Racial demographics could surface resulting in bias. For that reason, it was excluded in data preparation. Next it was discovered that the maximum loan amount allowed by the SBA is five million dollars (SBA, 2011). Since the SBA is responsible for up to 85% any loan, loans with an SBA approval of over $3,750,000 in the data set were invalid. Any loans over that would need to be removed in the data preparation phase.

During the data understanding phase, distribution was analyzed. This analysis was performed in R Studio. Most continuous variables indicated positive skew. This was backed up visually and by statistical methods. The following graph in figure two illustrates this finding.

Figure 2

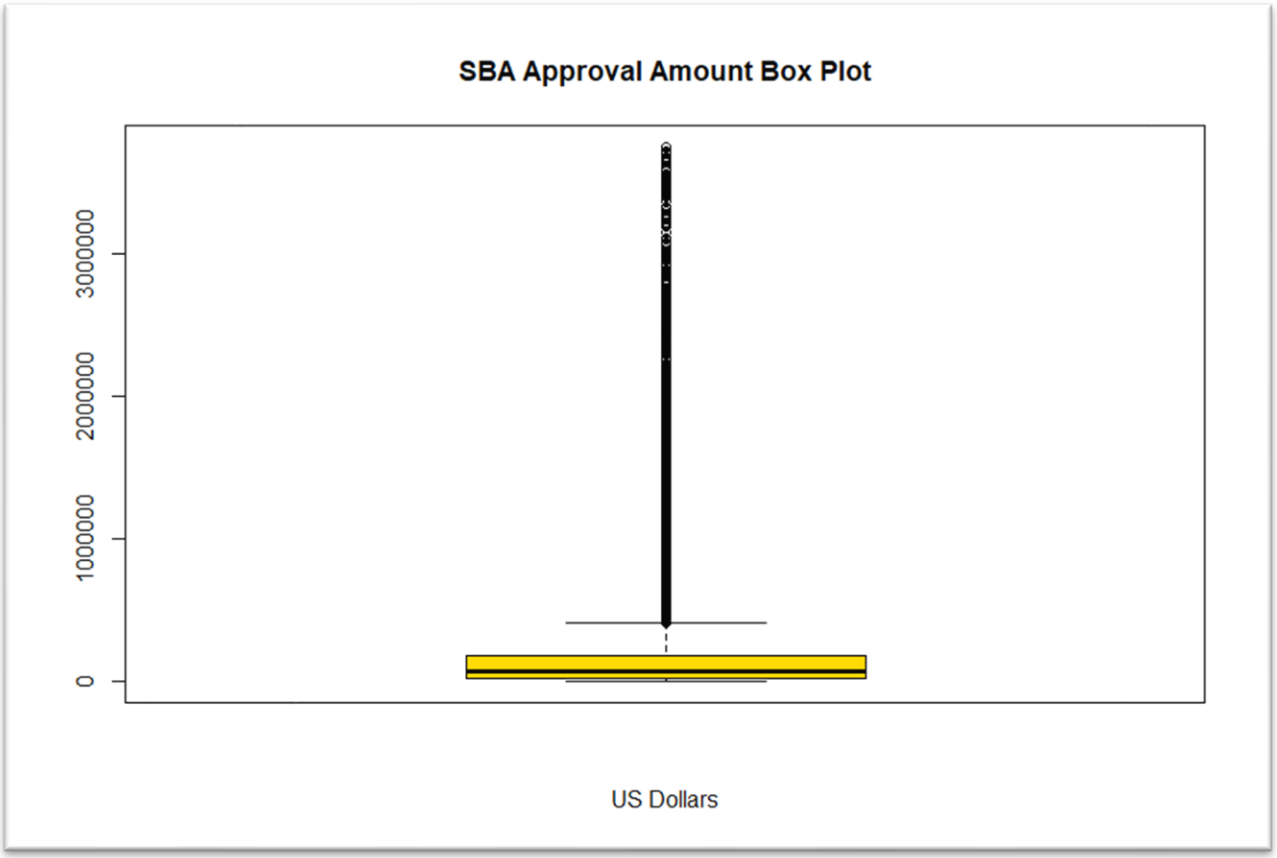
SBA Loan Approval Histogram



These findings of skew were typically exacerbated by many outliers. These were high positive values and were found on most of the continuous features. Except for invalid data mentioned in the business understanding phase, I decided to keep these outliers as they could have significant impact on predicting job creation values that were also highly affected by large positive outliers. Figure three is such an example.

Figure 3

SBA Approval Amount Box Plot



As such, Spearman's method was used to conduct correlation analysis. No significant correlation was noted for any feature to either target variable. To address the determination of a non-normal distribution, non-parametric models would be considered in addition to performing log ten transformations as necessary. Missing data was also evaluated, and it was determined that *chgoffdate* should be removed due to the large number of missing values in addition to the provisions against future data.

In the data preparation phase, several variables had to be converted to the appropriate data types. Encoding methods were deployed to address several categorical variables. One-hot encoding was the method of choice, but for features with high cardinality, binary encoding was also used. Rows and columns were removed in this phase as noted in the business and data

understanding phases. Two features were created. One was used to transform the information from *MIS_status* into a more usable format for modeling found in *MIS_logical.* The month a loan was approved was pulled from *approvaldate* and used to create the *month* feature which was later encoded via one-hot encoding. Features carrying risk of future data were removed. The data set was split into three sets for training, testing and eventual validation. This was done in a 60/20/20 ratio. The latter set was used to ensure against model overfit to the training and testing data sets.

Two models were required to best address the target variables. A classification model was used to determine if a loan would be repaid. Several models were compared using the Pycaret library. This library was also eventually used to tune the model with hyperparameters. Top contending models were ensemble models, and the Catboost model was selected as having the best fit for this data set. Since it was discovered in the analysis phase that most loans were repaid, a hyperparameter based on that ratio was provided to the model to allow for greater accuracy as measured by the area under the curve. Over forty features were eventually used to train the model. As a reminder, encoding leads to a splitting of one feature into several accounting for much of the increase in feature count noted from the original data set. Figure four shows all the models that were evaluated for the classification of loan repayment.

Figure 4

Loan Repayment Model Selection

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	CatBoost Classifier	0.9454	0.9769	0.9728	0.9614	0.9671	0.8075	0.8080	65.2207
1	Extreme Gradient Boosting	0.9437	0.9760	0.9718	0.9604	0.9660	0.8015	0.8020	35.0185
2	Light Gradient Boosting Machine	0.9399	0.9734	0.9724	0.9554	0.9638	0.7856	0.7867	3.1388
3	Gradient Boosting Classifier	0.9246	0.9580	0.9721	0.9386	0.9551	0.7217	0.7260	117.7144
4	Random Forest Classifier	0.9202	0.9241	0.9676	0.9376	0.9524	0.7077	0.7110	3.9060
5	Decision Tree Classifier	0.9118	0.8507	0.9450	0.9478	0.9464	0.6975	0.6975	7.9694
6	Ada Boost Classifier	0.9015	0.9360	0.9577	0.9254	0.9413	0.6369	0.6407	30.5169
7	Extra Trees Classifier	0.8971	0.9071	0.9716	0.9097	0.9396	0.5944	0.6095	48.8056
8	K Neighbors Classifier	0.8739	0.8320	0.9563	0.8974	0.9259	0.5053	0.5171	188.3537
9	Logistic Regression	0.8538	0.8432	0.9744	0.8652	0.9166	0.3429	0.3879	7.1251
10	Linear Discriminant Analysis	0.8410	0.8105	0.9807	0.8496	0.9104	0.2337	0.2958	4.8064
11	Ridge Classifier	0.8278	0.0000	0.9983	0.8281	0.9053	0.0443	0.1299	0.7120
12	Quadratic Discriminant Analysis	0.6430	0.5572	0.6935	0.8461	0.7569	0.0763	0.0838	2.4042
13	Naive Bayes	0.6114	0.7276	0.5781	0.9215	0.7061	0.2087	0.2657	0.6405

For predicting job creation, a regression model was deployed. Initially, this variable proved challenging to predict, but adding month and fiscal year from the approval features overcame that obstacle. First attempts were unacceptable and ultimately required the time-based features to achieve an acceptable level of performance. As in the previous model, the Pycaret library was used for model comparison and in parameter tuning. Ensemble models were the most effective in this evaluation as well. Light Boost Gradient Machine or LGBM model was eventually selected as having the best fit. Hyperparameters for the number of leaves and estimators proved the most beneficial in this case. Over fifty variables provided positive impact on the model and were ultimately selected to train the model. Figure five provides a list of regression models that were considered for predicting job creation.

Figure 5

Job Creation Regression Model Selection

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
0	Light Gradient Boosting Machine	2.5110	805.0595	24.0522	0.9840	0.8174	0.6456	2.6934
1	Random Forest	2.3437	907.0703	26.9691	0.9818	0.7628	0.8195	368.5672
2	CatBoost Regressor	2.5369	935.0904	27.3030	0.9814	0.8232	0.7145	34.7115
3	Extreme Gradient Boosting	2.5038	1016.1820	27.9235	0.9794	0.8106	0.7157	43.0497
4	Gradient Boosting Regressor	3.9209	1261.7134	33.5795	0.9750	0.9916	0.5829	133.1734
5	Decision Tree	2.7800	1465.9661	35.1829	0.9701	0.8811	1.2803	11.2244
6	Extra Trees Regressor	3.9291	3538.0158	57.0482	0.9304	0.8383	0.8744	432.3035
7	AdaBoost Regressor	17.6266	13920.7825	94.4928	0.6942	1.5104	1.8994	49.2989
8	K Neighbors Regressor	9.1642	29148.4861	170.2002	0.4267	0.9043	1.2477	47.4085
9	Ridge Regression	25.2379	50559.3897	224.1580	0.0135	2.4966	7.1524	0.3961
10	Linear Regression	25.2720	50563.8300	224.1669	0.0134	2.4975	7.1672	1.6443
11	Bayesian Ridge	24.9357	50569.8690	224.1805	0.0133	2.4800	7.0333	2.6291
12	Orthogonal Matching Pursuit	24.6727	50619.2977	224.2900	0.0123	2.4640	6.9642	0.5250
13	Lasso Regression	21.2700	50698.5964	224.4582	0.0109	2.3022	5.2321	1.4397
14	Elastic Net	18.9921	50798.7026	224.6749	0.0090	2.1926	3.4899	0.7443
15	Passive Aggressive Regressor	13.7915	51247.2672	225.6510	0.0005	1.7032	3.6709	1.4120
16	Lasso Least Angle Regression	13.4521	51270.4990	225.7075	-0.0000	1.9214	2.6455	0.5158
17	TheilSen Regressor	8.6419	51283.6966	225.7355	-0.0003	1.0095	0.7493	111.2804
18	Huber Regressor	7.9109	51332.1358	225.8422	-0.0012	0.9663	0.9624	47.0249
19	Least Angle Regression	24.6311	51699.5797	226.6965	-0.0098	2.4544	7.3779	0.5451

Results

The model for predicting loan repayment provided a 95% accuracy rating against the test data set and a 93% accuracy rating against the validation data set. The area under the curve for both showed a value of 93%.

The regression model returned a R-squared value of 99% for the testing data set. This indicates that 99% of the variance in the dependent variable is explainable by the model. Even when adjusted for the number of features used to train the model, a value of over 99% is returned. For the validation data set, both R-squared and adjusted R-squared values were 98%. The root mean squared error for the testing data was approximately 22 and for the validation data set was approximately 33. These values were high, but if taken in an aggregation could likely meet the original intent of maximizing the selection of job creating loans.

Discussions - Conclusions

Both models proved acceptable results at predicting their target variable. The classification model has proven the most successful and is production ready. Given the seasonality or time-based requirements for the regression model, I have reservations about recommending it for production without a warning of probable drift and ongoing support requirements. Skander Hannachi states that unlike traditional predictive models, time series forecasting requires model rebuilding before each prediction (2018). Even with constant retraining with the most current data, there is a two-year delay between when the SBA reports on jobs created and initial loan approval (SBA, 2010). It is unknown if this delay will prove problematic in a production setting. That is beyond the scope of my work. I would recommend any team implementing this model be made aware of these concerns and have a clear understanding of the objectives of both models. At a minimum, scheduling jobs that monitor drift is strongly recommended. Coding to support these models should be modular to support likely changes to address constant drift. Both models have been saved into pickle files for portability within the Python environment.

Acknowledgements

I would like to thank my family for their understanding of my time in this endeavor.

References

Hannachi, S. (2018, September 12). 3 facts about time series forecasting that surprise experienced machine learning practitioners. Retrieved from https://towardsdatascience.com/3-facts-about-time-series-forecasting-that-surprise-experienced-machine-learning-practitioners-69c18ee89387

Kagan, J. & Brock, T. (2020, May 19). Charge-Off. Retrieved from https://www.investopedia.com/terms/c/chargeoff.asp

Kaggle. (2020). SBA Loans Case Data Set. Retrieved from https://www.kaggle.com/larsen0966/sba-loans-case-data-set

Li, M., Mickel, A. & Taylor, S. (2018, April 5). Should This Loan be Approved or Denied? A Large Dataset with Class Assignment Guidelines. Journal of Statistics Education, 26(1), 55-66. https://doi.org/10.1080/10691898.2018.1434342

SBA. (2011, October). Loan Fact Sheet. Retrieved from https://www.sba.gov/sites/default/files/SDOLoanFactSheet_Oct_2011.pdf

SBA. (2010, April 10). ROM 10-15 – Review of SBA's Job Creation Data under the Recovery Act. Retrieved from https://www.sba.gov/document/report-10-15-rom-10-15-review-sbas-job-creation-data-under-recovery-act

Siegel, E. (2016). Predictive Analytics. Hoboken, NJ: Wiley.