**The World Bank IBRD Loan Data:**

**Analysis and Prediction Model of The Loan Cancellation**

**Kyoosik Kim**

**Abstract**

The International Bank for Reconstruction and Development (IBRD) has been in service since the group was formed in 1944. Moving into the modern era, the world largest development cooperative has provided financial services such as loans, risk management, and advising to middle and low-income countries [1]. Among these services, I focused on the loan service to analyze the recent trend of cancelled loans, eventually to build a predictive model as loan cancellations cause costs. I selected both the quantitative factors such as interest rates and loan amounts, and the qualitative factors such as regions and project sectors. I applied Mann-Whitney U test to the numerical data, and I used barplots for the categorical data. All the numerical variables appeared to be significant with Mann-Whitney U test, claiming that cancelled loans have different distributions from repaid loans. The categorical variables also showed differences as some of them divided cancelled loans from repaid loans. Through statistical methods like this, I selected several features to classify cancelled loans and built a prediction model using a decision tree which successfully classifies loans with virtually perfect accuracy.

**Introduction**

The loan service provided by the IBRD is nowadays considered as a financial solution which helps low-income countries with their economic development, although it was initially established for European countries after the World War II [1]. For this reason, I limited the period after 1980 based on the year each loan was closed. Considering these loans being a long-term, the recent four decades are old enough to have sufficient data and young enough to catch a modern trend. With these data, I first investigated the yearly trend to learn that loan cancellations occur when individual countries are struggling with economic recessions. Thus, my question became *“*What loans were cancelled during economic recessions?”. For the numerical features, I selected ‘interest rate’ and ‘original principal amount’ from the given columns, and I created ‘days from approval to signing’ and ‘days for repayment’. For the categorical features, ‘loan type’ and ‘region’ were selected and I created ‘project sector’ by clustering project names. My sub-question became “Are cancelled loans and repaid loans are statistically different?”, and Mann-Whitney U test and barplots were applied to answer it. Mann-Whitney U test is considered when a distribution does not follow the normal distribution [2], as do the numerical variables in this research.

**Methods**

The IBRD loan data is maintained by the World Bank from which I downloaded the data through their API service [3] with the help of *sodapy*, a Python package. I used other Python packages such as *pickle* for file management as I heavily used the Python for the entire project. Ahead of the analysis, I reduced the original data with 326k+ observations by filtering years of economic recession for each country, down to 41k+. The data originally consisted of 32 columns [3], however, I removed many of them as they were clearly in causal relationship with the dependent variable. For instance, the column ‘cancelled amount’ can only be effective when the loan was cancelled. I also removed all the date-related columns since my aim was not at a time series analysis, and I assume that the seasonality was mostly eliminated when I filtered the data by recessed years. However, I could manage to extract two useful numerical features from the date columns. Finally, I removed some categorical data which could be narrowed down to a single identical value. Then I processed the rest selected, especially, clustering hundreds of different project names shown on Figure 1 into countable project sectors using *fuzzywuzzy* package. Lastly, I used the balancing technique for classes [4] as repaid loans are dominant than cancelled loans.

Since the distributions are non-parametric [2], I conducted Mann-Whitney U test on the numerical data for which I used the function *mannwhitneyu* from *scipy.stats*. The null hypothesis for the test is that the two data are from the same distributions. I tested the data of repaid loans and the data of cancelled loans for each numerical feature. I used *seaborn.barplot* to draw the categorical features to compare how many loans were cancelled out of the total loans for possible values of each category. For example, I found out that countries in Africa cancelled half the loans while countries in East Asia and Pacific rarely cancelled the loans. I also utilized *seaborn.boxplot* to see if there is any relationship between numerical data and categorical data. From what I learnt through the analysis, I build a predictive model to predict a loan cancellation. Because all the features were significant to differentiate cancelled loans, I finalized the feature selection with the four numerical and three categorical data as shown on Table 1. I chose Naïve Bayes as the baseline classifier, and Support Vector Machine, Decision Tree, Gradient Boosting as the candidate classifiers [5]. Their performances were evaluated with accuracy, precision, recall, and f-score [6].

**Results**

The Mann-Whitney U test conducted on the numerical features resulted in significance, p-value < 0.05, to claim that the distribution of repaid loans and the distribution of cancelled loans are different. As the plotted distributions suggests on Figure 2, the difference is so distinctive that the p-values on Table 2 almost converge to zero. The mean interest rate for cancelled loans is found to be 5.55% which is lower than 6.17% of repaid loans, however, the median interest rate of cancelled loans is slightly higher; 6.93% versus 6.90%. But the statistics change substantially when many the loans with 0% interest rate were not considered, the mean for repaid being 6.42% compared with 6.85% for cancelled. For the mean and median, cancelled loans are smaller in ‘original principal amount’ than repaid loans; 25 million $ versus 20 million $, the medians for the two classes. The median ‘days for signing’ appears to be longer for cancelled loans by 9 more days while the mean is shorter. This is because there are quite many observations with value zero, as with ‘interest rate’. But ‘days for repayment’ is shorter both in the mean and median by far as shown on Table 2.

There are seven different values in ‘loan type’ on Figure 3, among which ‘Non Pool’ and ‘Pool Loan’ are two major types. ‘Bloan’ and ‘FSL’ experienced a 100% cancellation rate while ‘SCP EUR’ and ‘SCP USD’ never did. Along with the major two, ‘SNGL CRNCY’ shows a decent cancellation rate approximately ranged between 34.48%. Among the six regions on Figure 4, ‘Africa’ and ‘Europe and Central Asia’ have the highest cancellation rate, 50.06% and 39.58%, respectively. Out of the fifteen project sectors grouped by project keyword, ‘Agriculture’, ‘Emergency’, ‘Finance’, ‘Rehabilitation’ sectors are vulnerable to cancellation as seen on Figure 5.

The Naïve Bayes classifier, baseline predictive model, shows a poor performance of accuracy 0.54 (+/– 0.02) and f1-score 0.49 on Table 3. It does particularly a terrible job to recognize a repaid loan as such, securing only 0.22 recall for the class. However, the performance advances almost to a perfection with the Decision Tree classifier. It only fails to classify a few repaid loans as such out of over a thousand, and the SVC classifier almost never fails to correctly classify every class. Even though the Gradient Boosting classifier is considered more sophisticated than the others, it performs as accurately as 0.96 (+/– 0.01) with 0.97 f1-score. An interesting result is that a Decision Tree classifier with using most important two factors on Figure 6 can still score 0.89 accuracy and 0.88 f1-score.

**Discussion**

To answer the research question as well as predict loan cancellations, I selected and analyzed seven different variables. I discovered a tendency that cancelled loans generally have higher interest rates and they took longer to make the contracts effective than repaid loans. It is natural that a country would cancel loans with high interest rates when their economy is on a downturn. The reason why I suspect cancelled loans took longer time to sign on average could be due to qualitative factors such as ill-management or risks. For example, loans managed by an inexperienced person would take longer for processing, also likely to be susceptible to being made with mistakes. For another example, loans with high risk would take longer to review, but they should still pose more risks than others afterwards.

The loan type ‘B-Loan’ and ‘FSL’ are useful to recognize cancelled loans while ‘SCP USD’ does for cancelled loans. ‘B-Loan’ is a type of co-financing loan that two or more institutions join a single development project [7]. Unfortunately, it seems like this risk-sharing cooperation never worked out when each participant had to encounter only a partial damage and probably did not urge the loan accomplishment. ‘FSL’ stands for ‘Fixed Spread Loan’ where the spread means the gap between the ask price and the bid price in the FOREX market [8]. The ‘FSL’ in the IBRD data, though, is actually any loan based on either the fixed or variable spread, according to the IBRD data dictionary. In other words, these loans based on spread are currency sensitive unlike other types. For example, Mexico cancelled all the loans classified as ‘FSL’ type in 2008 when the USD and Peso exchange rate soared up [9] as shown on Figure 7. In summary, this type can be considered as prone to such financial variabilities that they could easily be cancelled. As for ‘region’ category, African countries are most likely to cancel loans probably due to their inadequate economic backgrounds. Following Africa, Europe and Central Asia are the regions with a high cancellation rate. The countries in these regions used to be the Soviet Union such as Poland, Latvia, or Kazakhstan. They all suffered from severe downfalls after the federation folded. These two regions account for small portion of the total, though, other regions showing a similar pattern. Lastly, ‘project sector’ does not show as strong patterns as the other two categorical features even if ‘Agriculture’ and ‘Finance’ sectors have a bit higher chances of cancellation than the other sectors.

Putting these all together, the Decision Tree classifier works incomparably better than the baseline classifier which is presumed to be due to the dependency assumption not being satisfied [10]. The SVC classifier is better than the Decision Tree classifier solely on performance view point. Nevertheless, the Decision Tree classifier should be chosen for the sake of the principle of parsimony, aka Occam’s razor [11]. Another model, the Gradient Boosting classifier, does not do as well as the others only to prove that a complicated model is necessarily better. Finally, one point I did not consider for the research is types of economic recessions. There are only a few or less years in which a country was in a trouble enough to cancel the loans, and these economic recessions should have come in different forms as the economy always does. Therefore, the future question becomes “What type of economic recession affects loan cancellations the most?”.

**References**

[1] “International Bank for Reconstruction and Development”, *The World Bank* [Online]. Available: <https://www.worldbank.org/en/who-we-are/ibrd>. [Accessed Nov 15, 2018].

[2] Nadim Nachar, “The Mann-Whitney U”, *The Quantitative Methods for Psychology*, 2008, vol. 4, No. 1, p. 13-20.

[3] “IBRD Statement Of Loans - Historical Data”, *Asian Development Bank* [Online]. Available: [https://dev.socrata.com/foundry/ finances.worldbank.org/x433-52wt](https://dev.socrata.com/foundry/%20finances.worldbank.org/x433-52wt). [Accessed Nov 15, 2018].

[4] “How to Handle Imbalanced Classes in Machine Learning”, *Elite Data Science*, 2017 [Online]. Available: <https://elitedatascience.com/imbalanced-classes>. [Accessed Nov 15, 2018].

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning, Stanford, 2013.

[6] Tavish Srivastava, “7 Important Model Evaluation Error Metrics”, *Analytics Vidhya*, 2016 [Online]. Available: <https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics>. [Accessed Nov 15, 2018].

[7] “Co-financing Partnership – What is Co-financing”, *Asian Development Bank* [Online]. Available: <https://www.adb.org/site/cofinancing/what-is-cofinancing>. [Accessed Nov 20, 2018].

[8] “Introduction to Trading - Floating and Fixed Spread”, *IFC Markets* [Online]. Available: https://www.ifcmarkets.com/en/introduction/floating-fixed-spread. [Accessed Nov 20, 2018].

[9] “US Dollar Peso Exchange Rate - Historical Chart”, *Macrotrends* [Online]. Available: <https://www.macrotrends.net/2559/us-dollar-mexican-peso-exchange-rate-historical-chart>. [Accessed Nov 20, 2018].

[10] Ahmad Ashari, Iman Paryudi, and A Min Tjoa, “Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool”, *IJACSA*, Vol. 4, No. 11, 2013.

[11] Phil Gibbs, “Occam’s razor”, *Department of Mathematics at University of California*, 1996 [Online]. Available: <http://math.ucr.edu/home/baez/physics/General/occam.html>. [Accessed Nov 20, 2018].

**Tables**

**Table 1. Features Used**

|  |  |  |
| --- | --- | --- |
| Feature | Data Type | Values |
| interest\_rate | Numerical | Mean: 6.04% (+/– 2.53) |
| original\_principal\_amount | Median: 2,330,000 $ |
| days\_for\_signing | Median: 37 days |
| days\_for\_repayment | Median: 5114 days |
| loan\_type | Categorical | B-Loan, FSL, Non-Pool, Pool-Loan,  SCP EUR, SCP USD, SNGL CRNCY |
| region | Africa, East Asia And Pacific,  Europe and Central Asia,  Latin America and North America,  Middle East and North Africa  South Asia |
| project\_sector | Agriculture, City, Construction, Education, Emergency, Environment, Finance, Health, Industry, Power, Rehab, Road, Transport, Water, Other |

**Table 2. Mann-Whitney U Test Results**

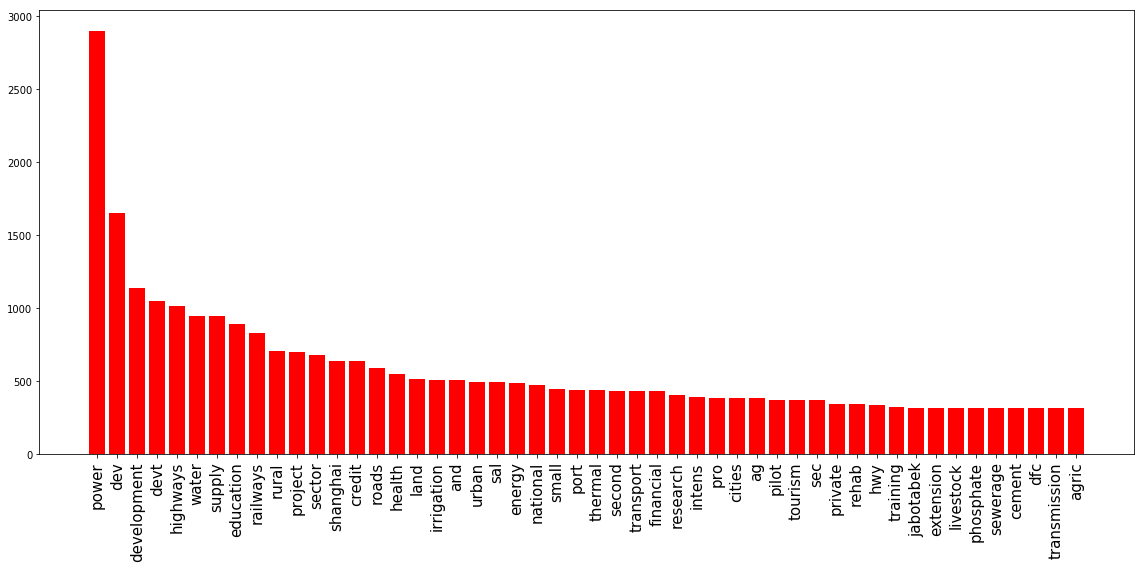
|  |  |  |  |
| --- | --- | --- | --- |
| Feature | P-value | Repaid  (mean / median) | Cancelled  (mean / median) |
| interest\_rate (%) | 0.00 | 6.17 / 6.90 | 5.55 / 6.93 |
| original\_principal\_amount (mil $) | 0.00 | 59.23 / 25.00 | 42.76 / 20.00 |
| days\_for\_signing (days) | 0.00 | 126.25 / 36.00 | 109.08 / 45.00 |
| days\_for\_repayment (days) | 0.00 | 4851.94 / 5294.00 | 4369.11 / 4564.00 |

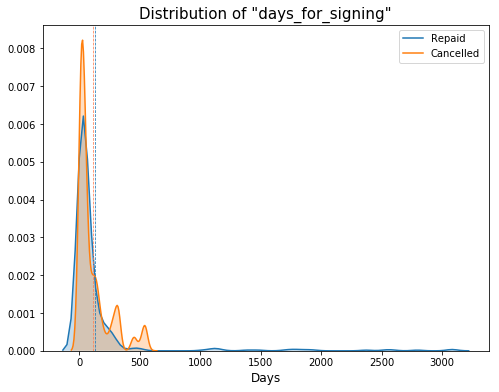
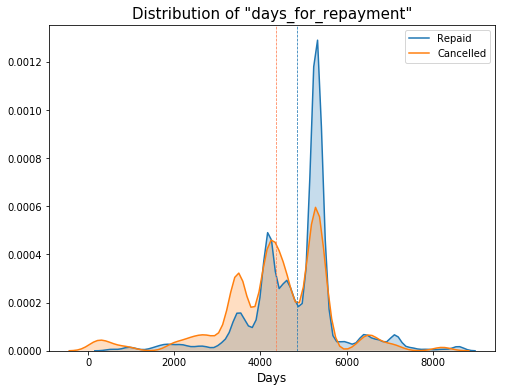
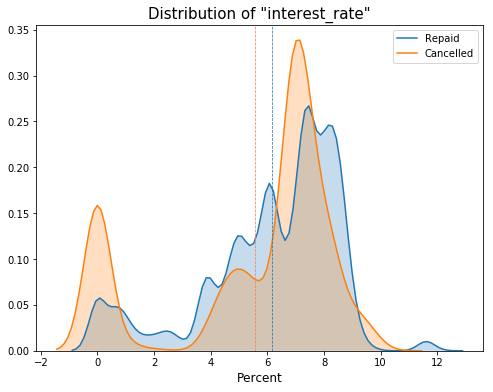
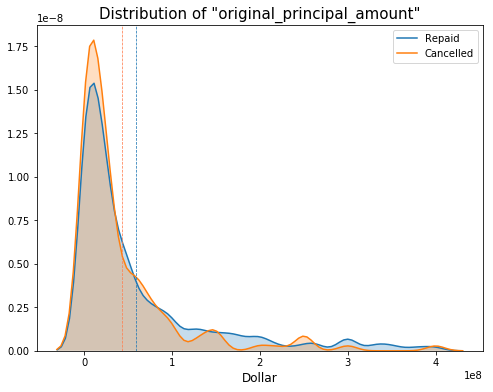
**Table 3. Model Performance Comparison**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score |
| Naïve Bayes | 0.54 (+/– 0.02) | 0.58 | 0.55 | 0.49 |
| Decision Tree (full) | 1.00 (+/– 0.00) | 1.00 | 1.00 | 1.00 |
| Decision Tree (partial) | 0.89 (+/– 0.01) | 0.88 | 0.88 | 0.88 |
| SVC | 1.00 (+/– 0.00) | 1.00 | 1.00 | 1.00 |
| Gradient Boosting | 0.96 (+/– 0.01) | 0.97 | 0.97 | 0.97 |

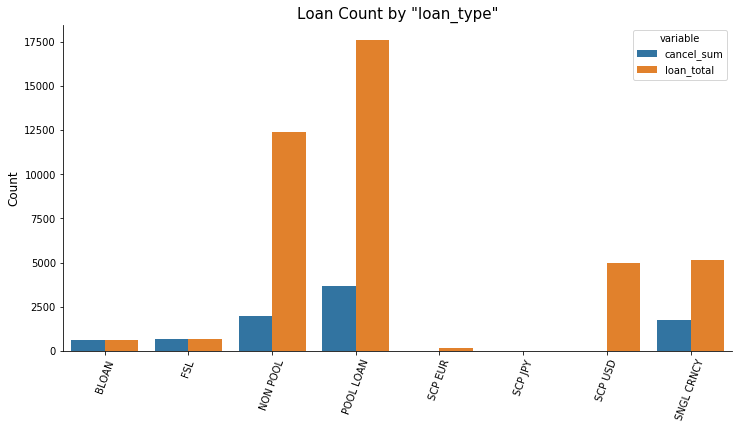
**Figures**

**Figure 1. Most Common Project Names**

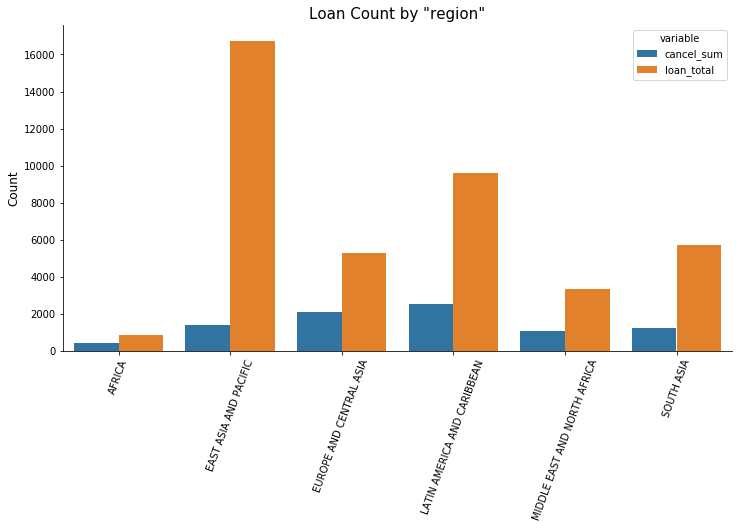
****

******Figure 2. Numerical Feature Distributions**

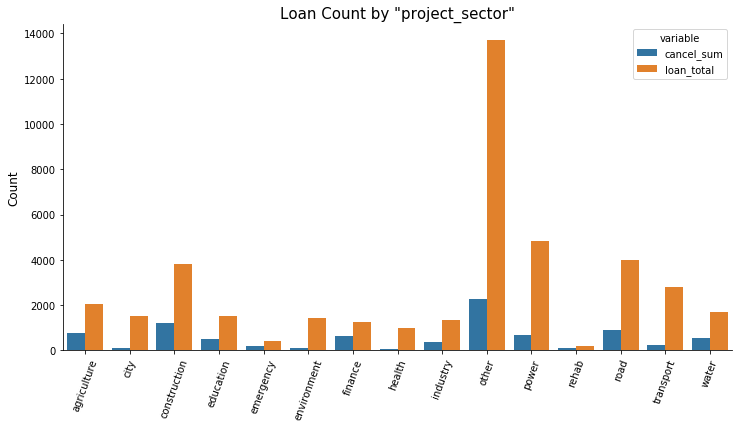
**Figure 3. Categorical Feature – “loan type”**



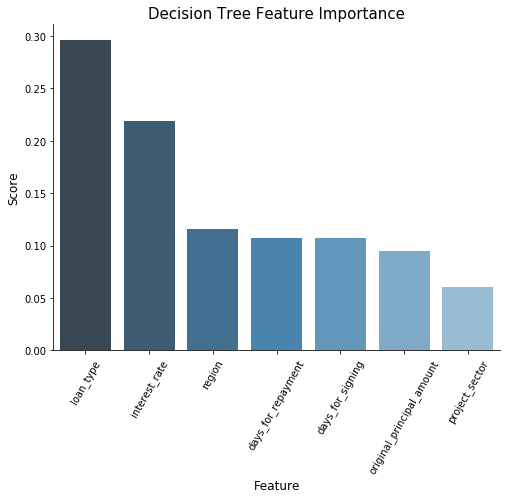
**Figure 4. Categorical Feature – “region”**



**Figure 5. Categorical Feature – “project sector”**



**Figure 6. Feature Importance**



**Figure 7. USD vs. Mexican Peso Exchange Rate**



**Supplemental Materials**

1. Project Python Code
2. IBRD Data Dictionary (pdf)
3. Decision Tree Visualization (pdf)