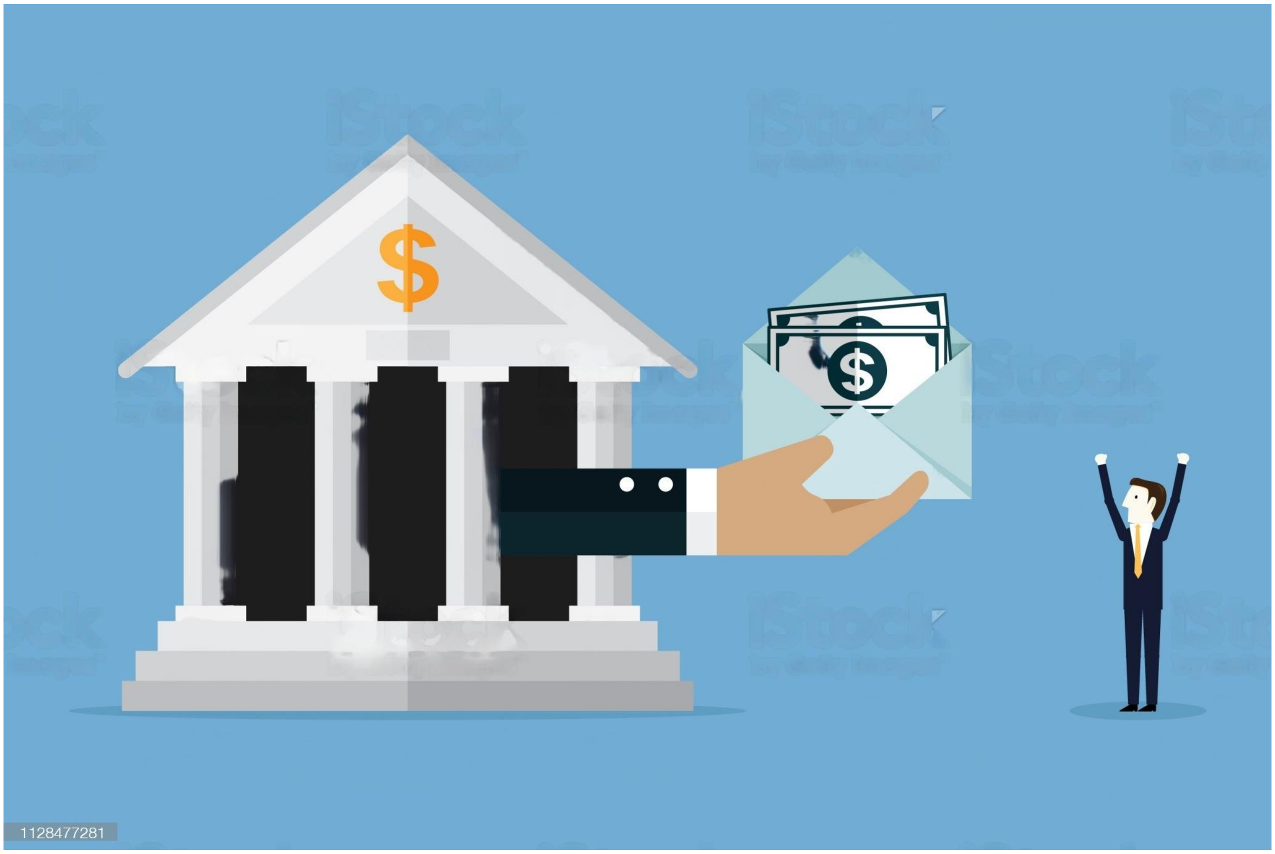

Analyse Risk on a Loan Portfolio using Statistical Tests in Jupyter

25.06.20

—

AMAN KUMAR
VIT Vellore

Overview

You are a data scientist for a finance company that relies heavily on data for decision-making in its day-to-day operations. After some evaluation, it has been found that the operational team uses raw data and their own assumptions to interpret the data, and this causes wrong decisions to be made. Recently, they assumed that the interest rates had no contribution to the default rates, so they approved loans as long as the customer income seemed promising. However, the risk team findings did not support this assumption.

Your role is critical in helping the company make data-driven decisions. Your manager tasked you with evaluating whether the operational team's assumption is right or wrong. As a data scientist, you must use the provided data to support your evaluation and use the proper statistical method to conclude.

Goals

1. Clean data using Jupyter Notebook
2. Calculate summary statistics such as mean and median from a data set.
3. Apply hypothesis testing and statistical tests to a data set
4. Interpret data results and decide whether an assumption about the data population is correct

Loading and Examining the Initial Data

Importing and exploring the dataset to see if the data is clean based on the below requirements:

- the data count is sufficient for analysis (at least 10.000 distinct rows)
- no blank data
- numeric values are in the acceptable range (e.g. loan interest percentage should not be negative or exceed 100%)
- the date range is a maximum of 12 months
- the loan channel contains one of the following valid strings: `DIRECT_SELLING`, `AGENT`, `WEB`, `MOBILE_APP`, `AFFILIATE`
- gender contains only `F` / `M` or `Female` / `Male`

Step-1+2: Data acquisition and exploration

```
Total rows: 15163

There are 163 duplicate rows in the Data Frame

There are NaN values in the Data Frame

“is_approved” True

“Is_default”

 True dtype: bool

principal_loan_amount: 1000 to 15000

request_date: 2023-01-01 to 2023-06-29

interest_rate: 7.5 to 10.0

loan_channel: ['WEB' 'MOBILE_APP' 'AGENT' 'AFFILIATE' 'DIRECT_SELLING'
'website_revamped' 'ANDROID_V2' 'ANDROID_V3' 'apple_new_v3']

is_approved: [True nan]

is_default: [nan True]

customer_monthly_income: 2000 to 6500

customer_age: 22 to 60

customer_gender: ['Male' 'Female' 'M' 'F']
```

Step 3:Data Cleaning

Duplicate rows dropped

Filled na values of “is_approved” and “is_default” with FALSE

Mapped “Customer_gender” such that only two gender categories are there: MALE=M and FEMALE=F

Mapped “loan_channel”.

Step 4: Hypothesis Testing

Evaluate The Impact of Interest Rates on Loan Default Rates

following hypotheses to test:

Null hypothesis (from the operations team) : the interest rate has no impact on the loan default rate

Alternative hypothesis: interest rate does have an impact on the loan default rate.

Used the two-tailed t-test to evaluate the impact of interest rates on loan default rates.

Splitting the data into two groups based on the interest rate levels, then comparing the mean default rates between the two groups to analyze the hypothesis.

Observation: The **t-statistic** is **-1.18** meaning that the average (mean) default rate for loans with high-interest rates is slightly lower than the mean default rate for loans with low-interest rates.

However, the **p-value** is **0.24**, which is higher than the commonly used threshold of 0.05 for statistical significance.

This means that the difference in default rates between the two groups may not be statistically significant. In other words, the interest rate has no significant impact on the loan default rate.

Step 5:Evaluate the Relationship Between Income Level and Loan Approval Rates

hypotheses to test:

Null hypothesis: customer income level has no impact on the loan approval rate

Alternative hypothesis: customer income level impacts the loan approval rate

Use logistic regression analysis to evaluate the impact of customer income level on loan approval rates. The customer income levels are separated into these groups:

- less than 2500
- 2500 to 5000
- 5000 to 7500
- more than 7500

Conclusion: The coefficient for the customer_monthly_income is -5.539e-06, which means that a one-unit increase in the customer_monthly_income results in a -5.539e-06 unit decrease in the log odds of loan approval (is_approved).

The p-value for this coefficient is 0.713, which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis that there is no linear relationship between income level and loan approval rates.

Step 6: Find Variable(s) That Affect Loan Default Rate

Since interest rate does not impact the default rate, you need to find out possible variable(s) that affect loan default rate.

Let's examine the following variable possibilities:

- customer_monthly_income
- customer_zip_code

we need to use regression test against data to find out

Conclusion: p-value for customer_monthly_income is greater than the significance level of 0.05, which means customer_monthly_income does not affect the default rate

p-value for customer_zip_code is less than the significance level of 0.05, which means customer_zip_code significantly affects the default rate

However, the warning message *Maximum Likelihood optimization failed to converge* suggests that the model may not have converged thus the results should be interpreted with caution

The warning *Possibly complete quasi-separation* indicates that there may be a perfect prediction or separation of the data points (in this case **customer_zip_code perfectly affects the default rate**), so the results should be interpreted with caution

Step 7:**Evaluate Loan Channel Against Default Rate**

additional hypotheses:

Null hypothesis: loan channel has no impact on the loan approval rate

Alternative hypothesis: loan channel impacts the loan approval rate

We will Use ANOVA test to determine if there is a significant difference between loan default rates across different loan channels. If the test indicates that there is a significant difference, you can use post-hoc tests to determine which specific loan channels have different loan default rates.

Conclusions:
Examine the post-hoc test.
If the reject column is true for a pair of loan channels, then you can reject the null hypothesis and conclude that there is a significant difference between the default rates of the loan channels.

In the post-hoc table, the reject column is true for the pairs of loan channels (AFFILIATE, WEB), (AGENT, WEB), (DIRECT_SELLING, WEB), and (MOBILE_APP, WEB). This suggests that the default rate for loans obtained through the **WEB** channel is significantly different from the default rates for loans obtained through the other channels.
Hence **we can reject the null hypothesis** since WEB channel contributes to loan default rate.

Possible reasons behind the impact of the WEB channel on higher loan default rates:

- **Customer Profile:** Evaluate the characteristics and demographics of customers who apply through the WEB channel to identify traits or behaviors associated with a higher likelihood of default.
- **Application Proces**s: Assess any differences in the application process for the WEB channel compared to other channels, such as the level of documentation required or the rigor of the approval process.
- **Customer Experience:** Examine the user experience within the WEB channel, including factors like user interface, transparency of information, and customer support, which could affect default rates.
- **Fraud Prevention:** Evaluate the effectiveness of fraud prevention measures specific to the WEB channel to ensure robust safeguards against fraudulent activities.
- **Risk Assessment:** Analyze the risk assessment methods used within the WEB channel and determine if any risk factors are not adequately considered, leading to higher default rates.

Recommendations for enhancing loan channel analysis:

- Conduct deeper investigations into the reasons behind the higher default rates in the WEB channel to identify specific factors contributing to the observed patterns.
- Evaluate customer profiles, application processes, customer experiences, fraud prevention measures, and risk assessment methods within the WEB channel.
- Use the findings to optimize loan channel strategies, improve risk assessment processes, and enhance the user experience within the WEB channel.
- Implement targeted measures to mitigate risks associated with the WEB channel, such as refining the risk assessment methodology and strengthening fraud prevention measures.
- Continuously monitor and assess the impact of the loan channel on default rates, ensuring that ongoing analysis informs decision-making and strategy development.