Risk Management

Credit Risk modeling with machine learning

Agenda

- Executive summary
- Risk management overview
- Project solution
- Credit Risk modeling
- Deliverables
- Next steps

Execute Summary

- This project aims to provide a simplified demonstration related to financial risk management, including credit risk modeling and the solution implementation process. The project will also showcase the data-driven calculation and simulation of expected loss (EL) and the probability of default (PD) using machine learning techniques.

- Credit risk modeling is the process of assessing the likelihood of a borrower defaulting on a loan or failing to repay debt. Credit risk modeling is an essential component of the lending process for banks and other financial institutions, as it helps to determine the creditworthiness of borrowers, the risk of default, and the appropriate level of interest rates to charge

- In recent years, there has been a growing interest in using machine learning and artificial intelligence (AI) techniques to improve credit risk modeling. These techniques can analyze large datasets and identify patterns that may not be evident in traditional statistical models, leading to more accurate risk assessments and better lending decisions.

Risk management overview

- Risk management is the process of identifying, assessing, and mitigating risks that may affect an organization's objectives. It is a critical part of any business strategy as it helps organizations to anticipate and prepare for potential risks, minimize their impact, and take advantage of potential opportunities.

- Effective risk management can help organizations to avoid or minimize potential losses, protect their reputation, and take advantage of new opportunities. It is an essential part of any organization's strategy for achieving its objectives and ensuring its long-term success.

- Types of Risk Management examples
  - Strategic risk management
  - Financial risk management
  - Compliance risk management
  - Project risk management

Credit risk is one of the key types of financial risk management

Financial risk management : involves the identification and control of financial risks such as credit risk, market risk, liquidity risk, and operational risk. The goal is to minimize the impact of these risks on a company's finances.

Project solution

- In this project, we have applied advanced data science techniques to develop highly accurate machine learning models that predict the probability of loan default. The solution package we provide includes Jupyter notebooks containing comprehensive details of the data preparation, feature engineering, and modeling techniques utilized. We also provide the trained machine learning models, dashboards, and other relevant artifacts that were instrumental in the development of the final solution.

- Data science plays a pivotal role in credit risk analysis, assisting lenders in gaining a better understanding of the risks associated with loan default. The use of advanced analytics and machine learning techniques enables lenders to make informed decisions based on highly accurate credit risk assessments, ultimately leading to better lending practices and reduced credit losses. One important metric used to assess credit risk is expected loss (EL), which represents the projected amount of loss that can be expected due to loan defaults. EL is calculated by multiplying the probability of default (PD) by the exposure at default (EAD) and the loss given default (LGD).

- By accurately predicting the probability of default using machine learning models, lenders can more effectively manage the risks associated with loan default and minimize their expected losses. Our project's solution package provides a comprehensive set of deliverables that can be used to calculate the expected loss for a given portfolio of loans. These deliverables include the trained machine learning models, dashboards for visualizing credit risk assessments, and other relevant artifacts that can be leveraged by lenders and other stakeholders to make better credit decisions and mitigate risks.

Credit risk modeling – Expected Loss (EL)

Quick review

Expected Loss (EL) is the amount of money a lender can expect to lose on average over the life of a loan due to default. It takes into account the probability of default, the exposure at default, and the loss given default. Here's how to calculate the Expected Loss:

```
EL = PD x LGD x EAD
```

Where:

- PD = Probability of Default: the likelihood that the borrower will default on the loan during the life of the loan.
  - `Machine Learning model (classification problem) with a PD of 10%`
- LGD = Loss Given Default: the amount of money the lender expects to lose if the borrower defaults on the loan.
  - `LDG = (Total exposure - Recoveries) / Total exposure = (USD 100,000 - USD 20,000) / USD 100,000 = 80\%`
- EAD = Exposure at Default: the amount of money the lender is exposed to when the borrower defaults on the loan.
  - `EAD = Total exposure x (1 - Recovery rate) = USD 100,000 x (1 - 0.20) = USD 80,000`
  - The recovery rate of current loan is going to be calculated with GBM model - recovery rate (regression problem)

To calculate the Expected Loss, you need to estimate each of these components based on historical data.

For example, suppose a lender has a USD 100,000 loan to a borrower with a probability of default of 10\%, a loss given default of 80\%, and an exposure at default of USD 80,000. The Expected Loss with formulas and number above, would be:

```
EL (result) = 10% x 80% x $80,000 = $6,400
```

This means that the lender can expect to lose $6,400 on average over the life of the loan due to default. The Expected Loss is an important metric for lenders because it helps them estimate the amount of risk they are taking on and set appropriate loan pricing and risk management strategies.

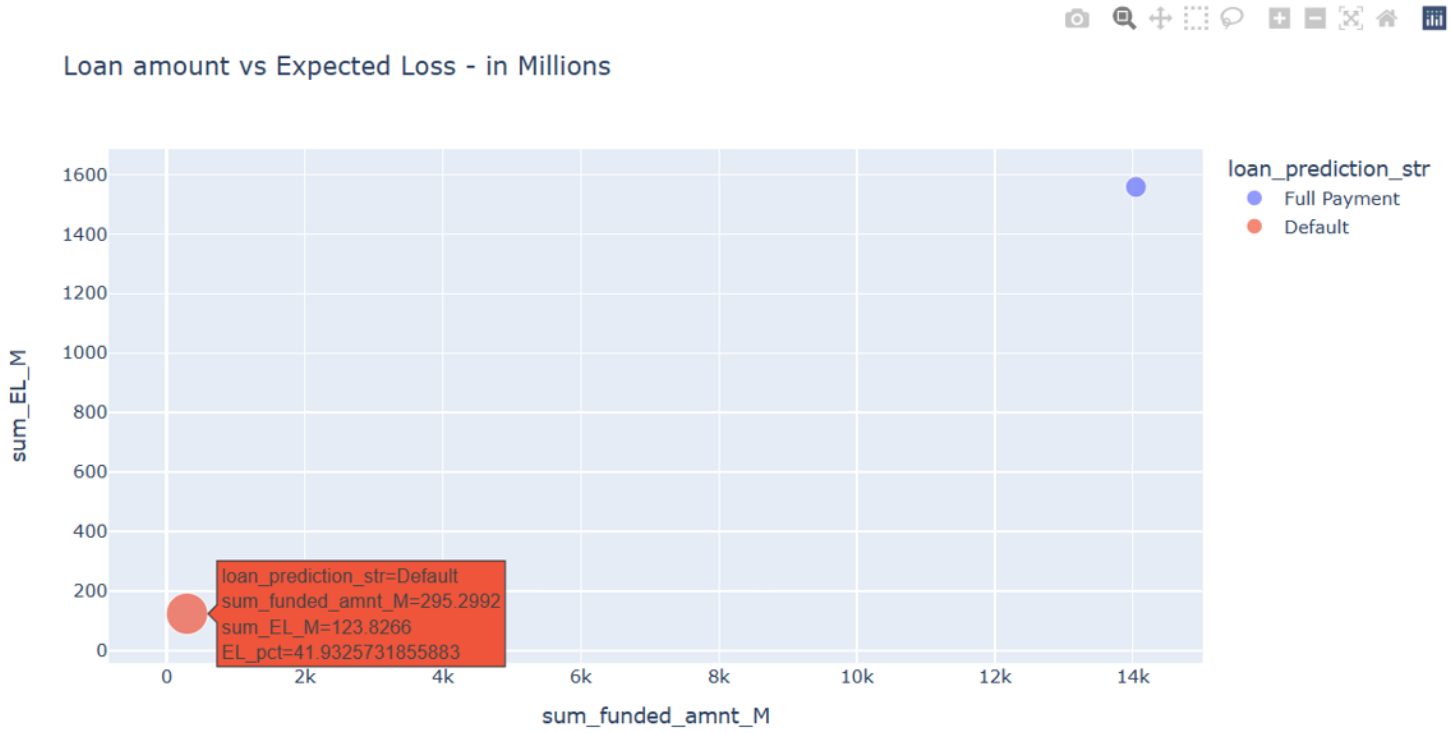
Notes

- Two machine learning models will be used to demonstrate how to approach this problem using historical data : one classification model to obtain the probability of default (PD) and one regression model to calculate the percentage of the recovery rate, which is used to calculate the EAD (exposure at default).

- The final EL (expected loss) is calculated by multiplying all the factors mentioned above.

Credit risk modeling – Expected Loss (EL)

- The expected loss percentage is much higher for loans predicted to default. It's important to note that this expected loss percentage takes into account factors such as the loan amount, interest rate, and probability of default. Lenders use this information to evaluate the risk of default and determine the appropriate interest rate to charge for the loan.

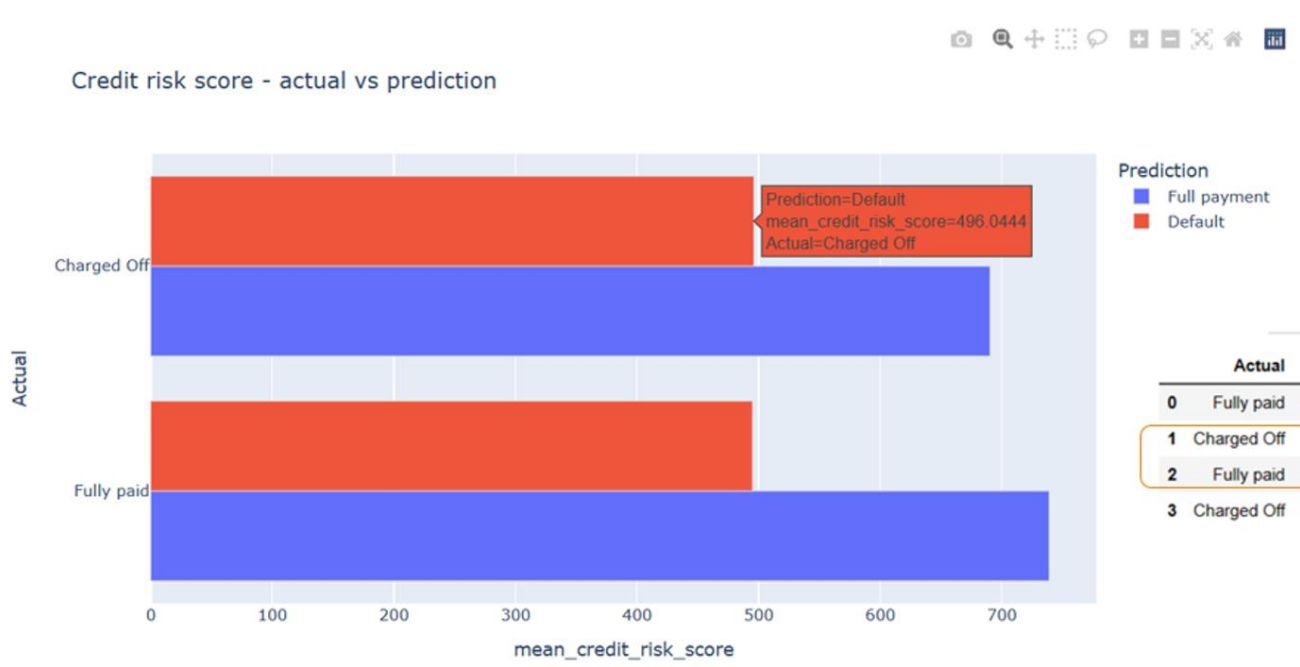


Expected Loss (EL): The expected percentage loss is quite high when compared to the full payment vs default payment forecast as can be seen in the size of the bubble sort. However, the final loss expectation is much lower when compared to full payment in the top right corner.

Credit Risk modeling – mean of credit risk score vs PD (probability of default)

- "Mean credit risk scores, which are used to assess the likelihood of default, were found to be lower for defaulting accounts. This suggests that accounts with lower scores had a higher probability of default. It's important to note that credit risk scores are based on a variety of factors, including credit history, income, and outstanding debts, and are used by lenders to evaluate the creditworthiness of borrowers.

- These numbers and considerations are also provided below as a result of the machine learning model to predict and build credit risk score



	Actual	Prediction	mean_credit_risk_score
0	Fully paid	Full payment	738.906921
1	Charged Off	Default	496.044353
2	Fully paid	Default	494.916955
3	Charged Off	Full payment	690.263199

Note: Results provided by the notebook solution using Spark and H2O Clusters at bin/ directory

Deliverables

Project Deliverables: Notebooks, Machine Learning Models, and Dashboards

- Notebooks containing comprehensive details of the data preparation, feature engineering, and modeling techniques with machine learning models
- Two machine learning models - one for classification and one for regression – for improved informed decision-making.
- Dashboard presentations to showcase project results and insights.

`EL = PD x LGD x EAD`

Where:

- PD = Probability of Default: the likelihood that the borrower will default on the loan during the life of the loan.
  - `Machine Learning model (classification problem) with a PD of 10%`
- LGD = Loss Given Default: the amount of money the lender expects to lose if the borrower defaults on the loan.
  - `LDG = (Total exposure - Recoveries) / Total exposure = (USD 100,000 - USD 20,000) / USD 100,000 = 80\%`
- EAD = Exposure at Default: the amount of money the lender is exposed to when the borrower defaults on the loan.
  - `EAD = Total exposure x (1 - Recovery rate) = USD 100,000 x (1 - 0.20) = USD 80,000`
  - The recovery rate of current loan is going to be calculated with GBM model - recovery rate (regression problem)

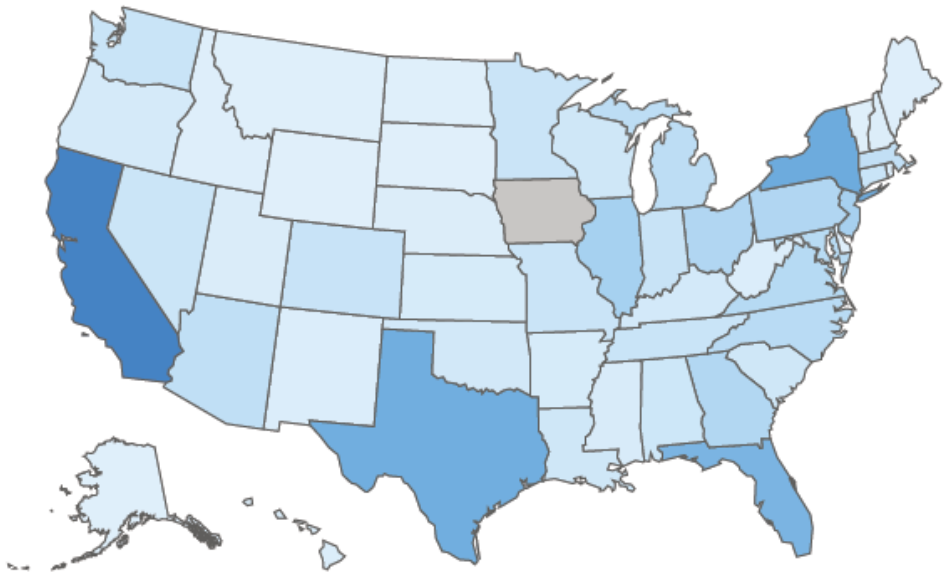
To calculate the Expected Loss, you need to estimate each of these components based on historical data.

Sample Dashboard

Expected loss by state in 2018

906,2K	$1.855,4	12,72%	$454,7	21,97%
number_of_contracts	Average of EL	Average of interest_rate_pct	Average of installment	Average of PD

Sum of EL by addr_state



addr_state	number_of_contracts
CA	117196
TX	75525
NY	75318
FL	65557
IL	39135
NJ	34178
OH	31114
PA	30947
GA	30448
NC	24720
VA	24578
MI	23333
MD	22291
AZ	20973
MA	20559
CO	18477
WA	17804
CT	15836
Total	**906193**

Dashboard | Credit_Risk_EL | Influencers_EL | Interest_rate | EL_by_State | EL_by_State_TX | Loan_contract_ID | EL_calculation

Next steps

After successfully completing Phase 1 of the project, the next steps could be:

- **Phase 2:** Full automation of the data pipeline and model prediction to optimize and ensure smooth and efficient data processing.

- **Phase 3:** Improving the machine learning models used for decision-making. This will involve incorporating new features and additional feature engineering steps, clustering loan transactions to identify patterns and trends, and executing A/B tests to verify the effectiveness of the improved models.

- **Phase 4:** Integration of the entire solution using the Data Lakehouse architecture. This will ensure the seamless integration of all project components, including the data pipeline and machine learning models, into a unified solution that delivers accurate insights and informed decisions related to risk management.

The chart below shows a summary of the steps involved in the data pipeline process