Slide 1: Thank you for joining me, fellow data scientists. I’m here to introduce an innovative way to assess bank loan portfolio risk using machine learning

Slide 2: Why Machine Learning?

Slide 3: Some benefits include but are not limited to: automation – reduces time spent on analysis and minimizes errors, accuracy – this allows the banks to take preventive measures, leading to a reduction in loan loss risk, cutting-edge technology - outperform competitors who are not implementing ML by increasing their overall portfolio

Slide 4: Now let’s explore the prediction pipeline using machine learning

Slide 5: The pipeline is quite simple, I input the loan portfolio in the classification model, which outputs a group of risky loans. Then, that dataframe of risky loans is used as input for the Linear Regression model. The final product is a predicted loan loss amount

Slide 6: Let’s discuss the data!

Slide 7: The data is a loan portfolio of about 100 thousand loans. The features are 770 anonymized, numerical columns. The target is the loan loss, based on percentage of loan lost. For example, a loan loss of 25% means 75% of the loan was paid back, and 25% was “lost” through charge-offs, bankruptcy, etc. And as you can see from the distribution, the target is highly imbalanced.

Slide 8: Methodology

Slide 9: First step was EDA - this was pretty straightforward. Second step was model selection – I used undersampling to fix the imbalance problem, hyperparameter tuning to optimize parameters, and feature engineering to improve accuracy. The final product was 2 XGBoost models, which I used for both classification and regression.

Slide 10: Drum roll please…

Slide 11: The classification model has a strong recall score and shows a high ROC AUC. Precision is low, but that’s not a problem. Recall is far important because it ensures potential losses do not fall through the cracks. The regression model shows a strong ability to predict the loan loss with minimal error.

Slide 12: The models predicted nearly 6,400 loans, which makes up about 6% of the total portfolio. And the average loan loss was about 1.6%, with the minimum being 0.43% and the highest predicted being about 35%. These numbers are significantly different compared to the actual numbers. This is likely due to the large number false positives, as implied by low precision.

Slide 13: The probability density functions of the actual and predicted loan loss also vary significantly.

Slide 14: For the next steps…

Slide 15: …I plan on researching why exactly my model’s predictions vary significantly from the actuals. Correcting this error will yield an even better prediction pipeline.