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 oversampling to fix the imbalance problem, hyperparameter tuning to optimize parameters, and feature engineering 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 does moderately well and scores a high AUC ROC. Recall was chosen because it was important to ensure 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 2,500 loans, which makes up about 2.35% of the total portfolio. And the average loan loss was about 3%, with the minimum being 0.44% and the highest predicted being about 25%

Slide 13: The distributions of the actual and predicted loan loss are similar.

Slide 14: For the next steps…

Slide 15: …I plan on researching why there’s a difference in scale between my predicted loss distribution and the actual loss distribution. My model is not able to predict large loan losses, which can cause several issues. Correcting this error will yield an even better prediction pipeline.