Capstone Project: Loan Default Prediction

Final Report

Matthew Wolf

Executive Summary

- Background: banks want to minimize loan defaults and classification models can aid in determining which potential applicants should get a loan
- **Objective: build classification models to predict loan defaults using the provided dataset, determine which model should be used, and provide recommendations to the bank based on important data features**

- Different data features were explored and visualized
  - Debt-to-income ratio, number of derogatory reports, and number of delinquent credit lines correlated with defaults
- Data was treated using different methods
  - Missing values were filled with median or mode, outliers were either removed or left in dataset
  - Imbalance of outcome variable was addressed using SMOTE analysis

Executive Summary

- Data was separated into categorical dependent variable ('BAD') and independent variables and data was split into training and testing sets 70% to 30%
- Different models were explored:
  - Decision Trees: weights/no weights, 75/25 and 80/20 data splitting, SMOTE analysis, tuning with different hyperparameters
  - Random Forests: weights/no weights, 75/25 and 80/20 data splitting, SMOTE analysis, tuning with different hyperparameters
  - Logistic Regression: with and without Lasso regularization, SMOTE analysis, threshold tuning
  - K Nearest Neighbors: SMOTE analysis, determining K value, tuning with hyperparameters
- Recall, precision, and accuracy were the metrics used to evaluate different models, especially recall to minimize false negatives (applicants who default on loans, but are predicted to repay)
- Tuned decision tree and feature importances were plotted

Executive Summary

- **Key takeaways**:
  - Leaving in outliers resulted in greater recall and accuracy as opposed to treating outliers
  - Generally, balancing the data with SMOTE analysis resulted in higher recall and accuracy
  - After tuning, decision tree, random forest, and KNN models were all able to achieve greater than 70% recall
  - The tuned KNN model resulted in the highest recall of 80%, with 95% accuracy and 97% precision on the test data
  - Debt-to-income ratio and number of delinquent credit lines were the most important features for predicting defaults; age of oldest credit line, home value, and number of derogatory reports were also important
- **Next Steps**:
  - The KNN model could be tuned again to further reduce overfitting the training data
  - To improve model further, more data could be collected, especially for debt-to-income ratio (which has 21% null values in initial dataset)
  - Build a data pipeline to use model with new applicants

Problem and Solution Summary

- **Problem**:
  - Banks want to minimize loan defaults, as defaults can result in major profit losses
  - Manual reviews of clients can be time-consuming and subject to reviewer biases
  - Machine learning models can be built using recent data to automate the process and try and remove biases
  - Models should try to minimize false negatives and maximize recall

- **Solution Design**:
  - Investigate different data treatment methods and different models and evaluate with recall, precision, and accuracy metrics
  - Compare different models on the basis of these metrics, and on the interpretability and implementation of these models
  - Make recommendations based on these models

Problem and Solution Summary

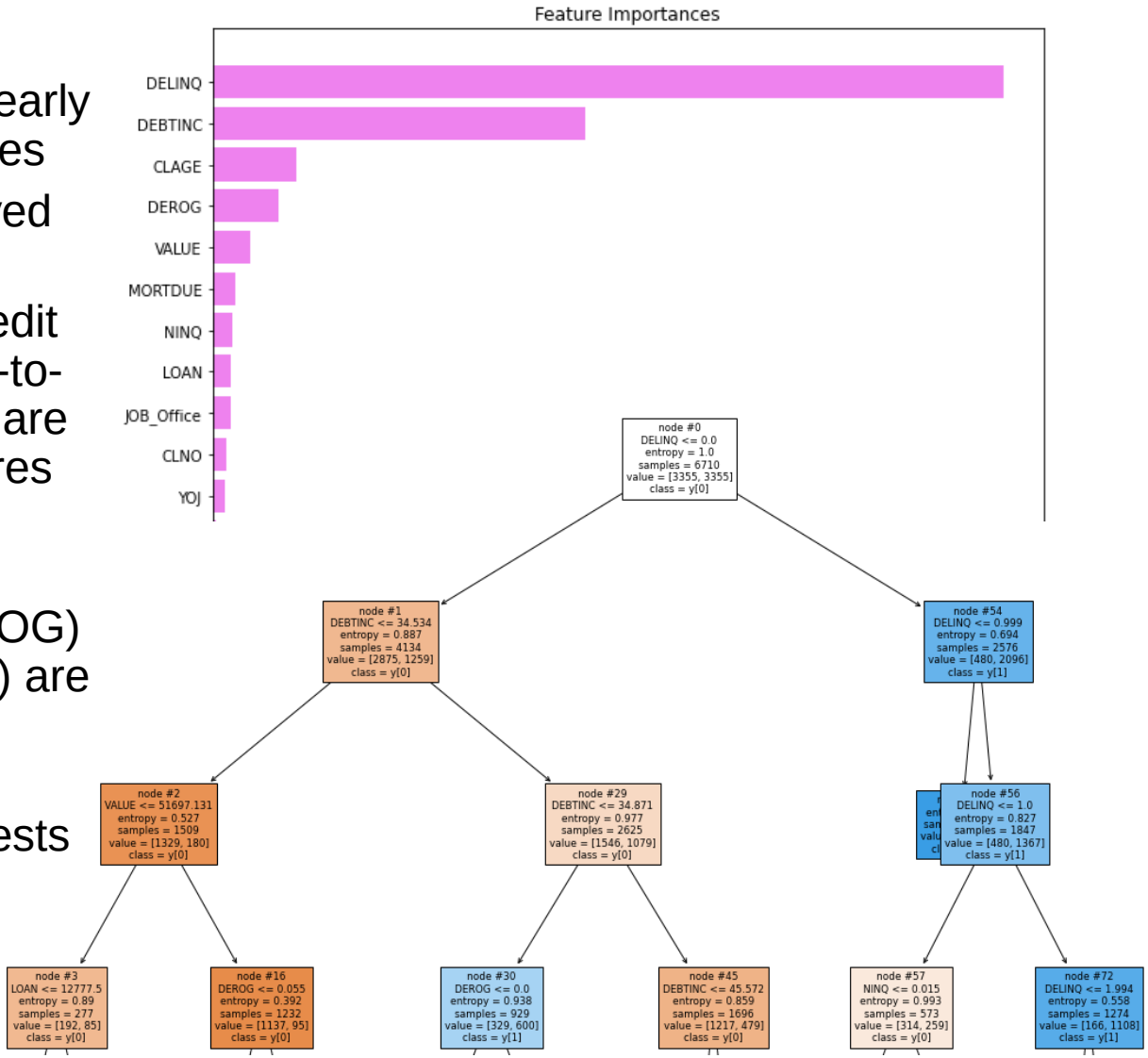
- **Model Comparison:**

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision
K Nearest Neighbors Tuned (k = 3)	1.00	0.95	1.00	**0.80**	1.00	0.97
Random Forest Initial	1.00	0.94	1.00	**0.77**	1.00	0.92
Decision Tree Tuned (6 Hyp)	0.85	0.84	0.83	**0.77**	0.85	0.58
K Nearest Neighbors	0.94	0.88	0.93	**0.72**	0.94	0.69
Random Forest Tuned (6 Hyp)	0.88	0.87	0.85	**0.71**	0.91	0.69
Logistic Regression	0.72	0.75	0.67	**0.67**	0.74	0.43
Logistic Regression with Lasso	0.72	0.75	0.67	**0.67**	0.74	0.43
Decision Tree Initial	1.00	0.87	1.00	**0.64**	1.00	0.69
Decision Tree Tuned (3 Hyp)	0.90	0.87	0.87	**0.61**	0.93	0.70
Random Forest Tuned (3 Hyp)	0.91	0.87	0.88	**0.60**	0.94	0.71

- Data was treated by filling null values, keeping outliers, splitting data 70/30 into train/test sets and balancing train data with SMOTE
- After tuning, decision tree, random forest and KNN models all had greater than 70% recall
- Tuned KNN had highest test recall, accuracy, and precision despite overfitting the training data
- Tuning sometimes resulted in lower recall when using SMOTE datasets

Problem and Solution Summary

- **Feature Importance**
  - Important features are clearly displayed by decision trees
  - The decision tree displayed below is cropped
  - Number of delinquent credit lines (DELINQ) and debt-to-income ratio (DEBTINC) are the most important features
  - Age of oldest credit line (CLAGE), number of derogatory reports (DEROG) and home value (VALUE) are also important
  - These features are also important for random forests and logistic regression



Recommendations for Implementation

- The tuned K nearest neighbors resulted in the highest test recall: 80% alongside 95% accuracy for the test data
  - Although this model overfit the training data to some extent, it still performed well on the test data and could be used with incoming loan applicants
- Through model development and visualization, some key features of the data that affect loan defaults/repayment have been determined
  - **Number of delinquent credit lines**: past delinquency on credit was an important model feature leading to more defaults
  - **Debt-to-income ratio**: a higher debt-to-income ratio was used to predict a higher likelihood of defaulting
  - **Age of oldest credit line**: applicants with older credit lines were more likely to repay their loans
  - **Other important features**: number of derogatory reports, value of home property, number of recent credit inquiries

Recommendations for Implementation

- Key recommendations:
  - Use tuned KNN model for future loan applicants
  - Build and maintain data pipeline to allow for efficient use of model with new applicants
  - Make sure stakeholders understand what information should be collected about new applicants to input into the model

- Key actionables for stakeholders:
  - For future applicants, collect information on debt-to-income ratio and delinquent credit history, as there were many missing values for both of these features
  - If possible collect other information about applicants such as married/family status, education history, while carefully considering possible biases that this data may introduce
  - Maintain consistent data pipeline and records for when model is used and with which applicants

Recommendations for Implementation

- Benefits and Costs:
  - **Benefits**: KNN model has high accuracy, is relatively easy to understand (grouping entries by other similar entries), once tuned is not too computationally expensive
  - **Costs**: model is less interpretable compared to some other models such as decision trees, which is why comparison to other models is necessary to extract important features

- Risks and Challenges:
  - Debt-to-income ratio was one of the most important features, but there was 21% null values for this feature
  - The model may be turning away applicants with a high debt-to-income ratio who are looking to consolidate debt and get back on their feet
  - Some younger applicants may not have much of a history of credit, debt, or employment and it may be difficult for the model to successfully gauge whether or not they will repay their loans

Recommendations for Implementation

- Further Analysis/Work:
  - Continue to try and tune models with more hyperparameters to reduce overfitting
  - SMOTE balanced datasets seemed to lead to greater overfitting during tuning
  - If possible, collect more data on recent applicants, especially debt-to-income ratio to improve model
  - Test other data treatment methods with KNN model, other models
  - Possibly fewer features (removing ‘REASON’ and ‘JOB’) could result in a more generalizable model