Automating Deferral Discrepancy Report

Springboard Data Science - Career Track
Capstone Three - Final Presentation
Marti Williams Kenna

Background

Problem:

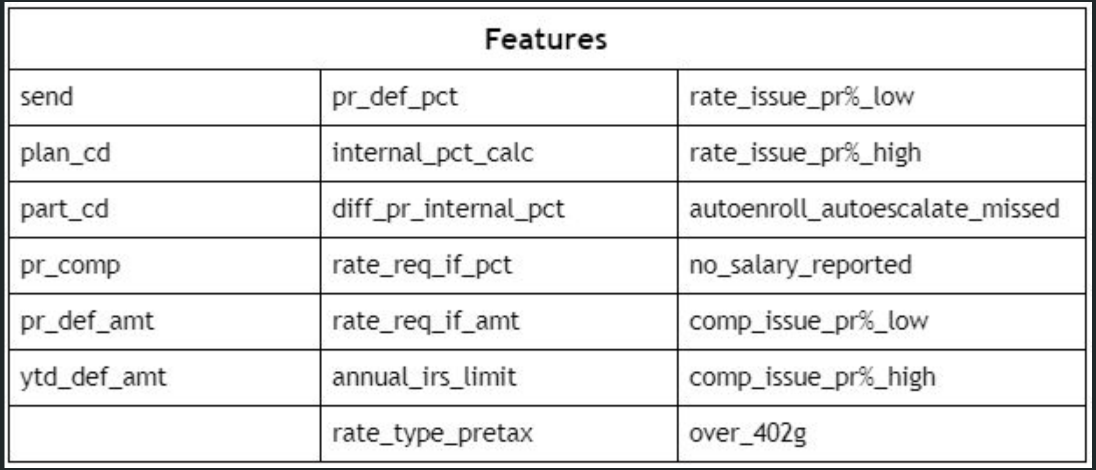
Can true discrepancies in participant deferrals be accurately identified using a supervised learning classification model?

IRS

- Business Analyst at Company B
- 401k plan management
- Why do we care?

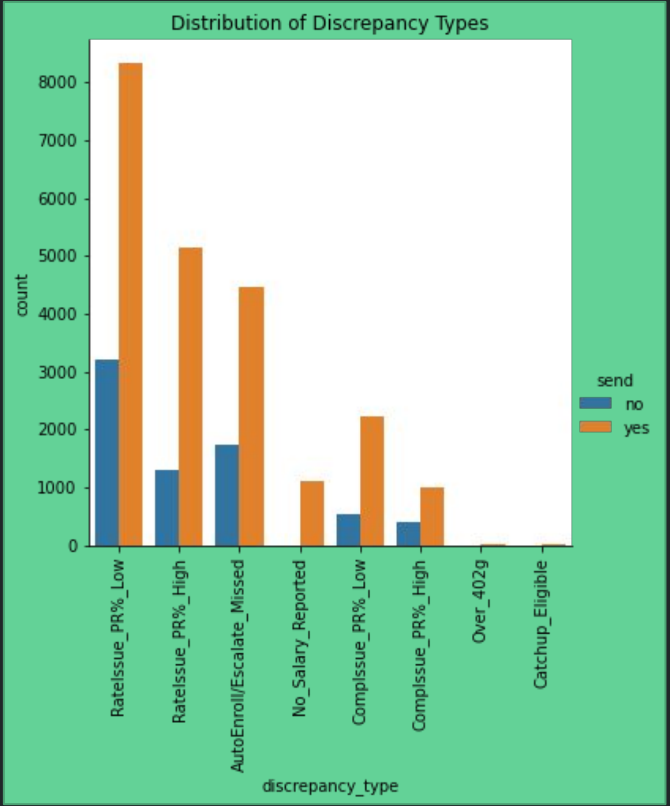
Data Collection

- Records from 1/1/2021 to 6/1/2021 (approx. 30000 records)
- Features included:

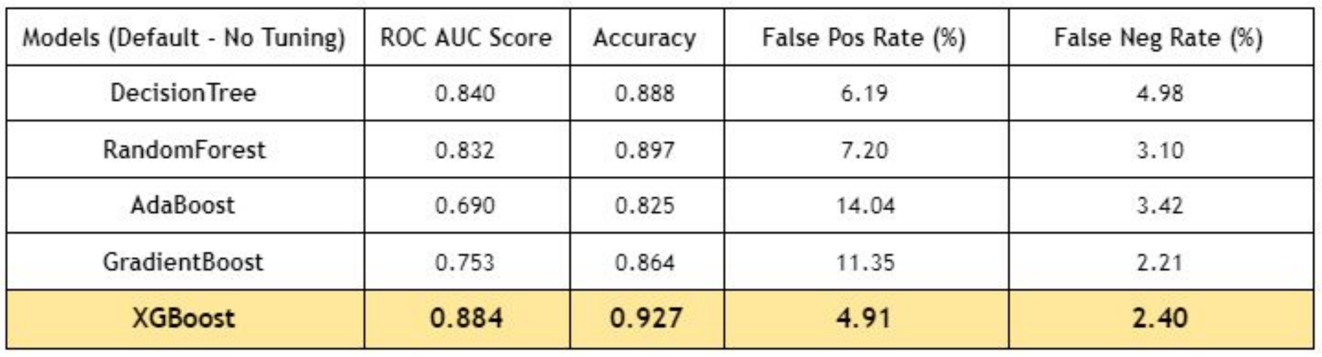


Features		
send	pr_def_pct	rate_issue_pr%_low
plan_cd	internal_pct_calc	rate_issue_pr%_high
part_cd	diff_pr_internal_pct	autoenroll_autoescalate_missed
pr_comp	rate_req_if_pct	no_salary_reported
pr_def_amt	rate_req_if_amt	comp_issue_pr%_low
ytd_def_amt	annual_irs_limit	comp_issue_pr%_high
	rate_type_pretax	over_402g

Data Insights

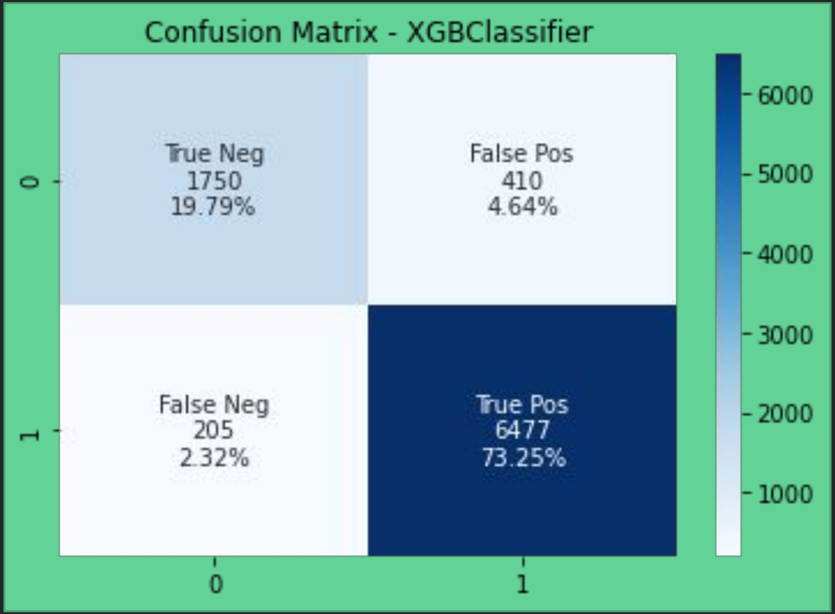


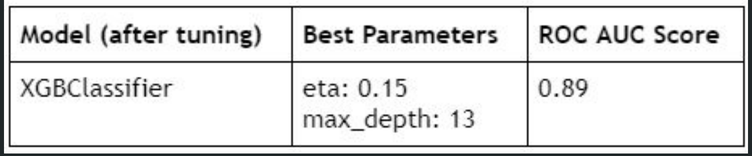
Model - Testing



Models (Default - No Tuning)	ROC AUC Score	Accuracy	False Pos Rate (%)	False Neg Rate (%)
DecisionTree	0.840	0.888	6.19	4.98
RandomForest	0.832	0.897	7.20	3.10
AdaBoost	0.690	0.825	14.04	3.42
GradientBoost	0.753	0.864	11.35	2.21
XGBoost	**0.884**	**0.927**	**4.91**	**2.40**

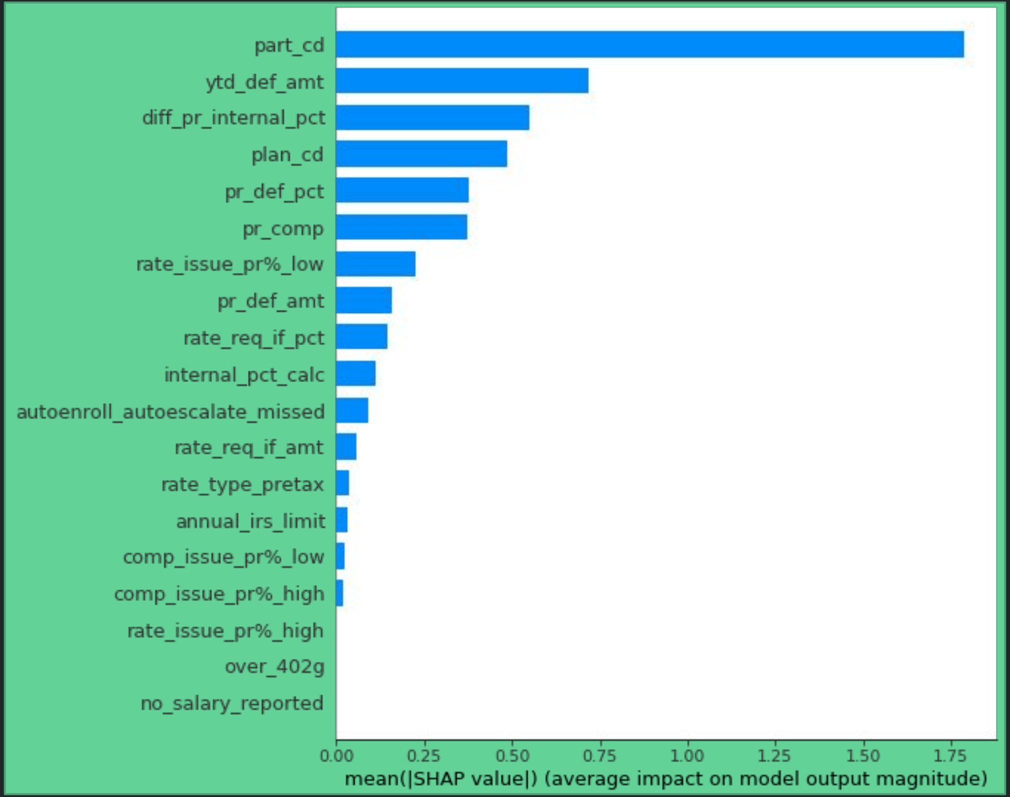
Model - Selection





Model (after tuning)	Best Parameters	ROC AUC Score
XGBClassifier	eta: 0.15 max_depth: 13	0.89

Model - Feature Importance



Future Scope/Conclusion

- Multi-class problem
- Expanded data set (2020 data)
- Fully automated process

Questions