Assignment 3 - Machine Learning
Workforce Retention Analysis

Mohammad Hossein Basouli

June 8, 2025

1 Introduction

In this analysis, we try to predict whether an employee leaves the company, based on some given attributes or not. *Institutional Attributes.*

2 Data

The training dataset contains 1341 rows and 35 columns. Features include information about employee's tenure time and duration, demographic, education, satisfaction & engagement factors and expertise.

3 Exploratory Data Analysis

- **Data Imbalance**: As we can see in the figure 1, the data is highly imbalanced between the two classes; only about %15 of the data is in the class 1 and the rest lies in the other class.

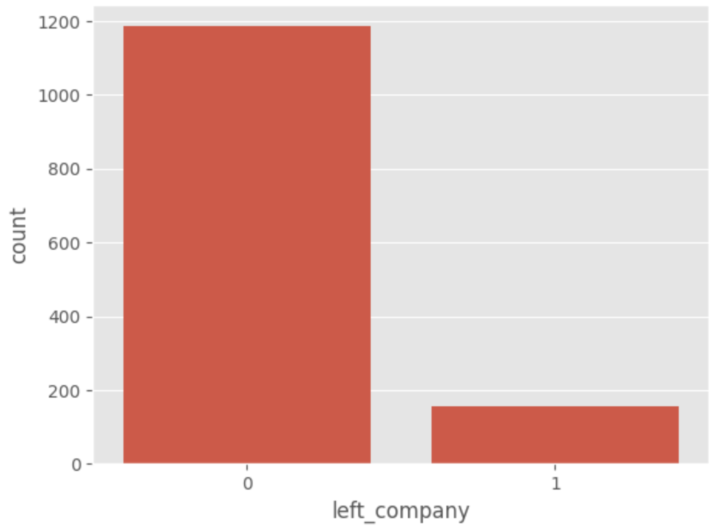

Figure 1: Data Imbalance: %15 Class 1 - %85 Class 0

- **Redundant Features**: By looking at the figure 2, we figure out that the feature *is_adult* is redundant.

```
Unique Values for travel_freq:
travel_freq
rare_travel        1027
frequent_travel     210
no_travel           104
Name: count, dtype: int64
----------------------------
Unique Values for work_division:
work_division
rnd      926
sales    384
hr        31
Name: count, dtype: int64
----------------------------
Unique Values for degree_field:
degree_field
life_sci     618
medical      436
marketing    123
tech_deg      95
other         57
hr            12
Name: count, dtype: int64
----------------------------
Unique Values for sex:
sex
male      849
female    492
Name: count, dtype: int64
----------------------------
Unique Values for job_title:
job_title
sales_exec      291
research_sci    276
lab_tech        256
mfg_dir         164
health_rep      123
manager          83
sales_rep        63
research_dir     59
hr               26
Name: count, dtype: int64
----------------------------
Unique Values for marital_state:
marital_state
married     621
single      451
divorced    269
Name: count, dtype: int64
----------------------------
Unique Values for is_adult:
is_adult
yes    1341
Name: count, dtype: int64
----------------------------
Unique Values for overtime_status:
overtime_status
no     1032
yes     309
Name: count, dtype: int64
```

Figure 2: Redundant Features: the column *is_adult* takes on a single value, *yes*

- **Correlated Features**: The heatmap in the figure 3 shows that some of the features, such as *years_with_manager* and *tenure_years* are highly correlated, thus causing a potential multi-colinearity. But as we have exprimented with replacing them by a single one of them, it yielded poor prediction results, thus we decided not to touch these features.

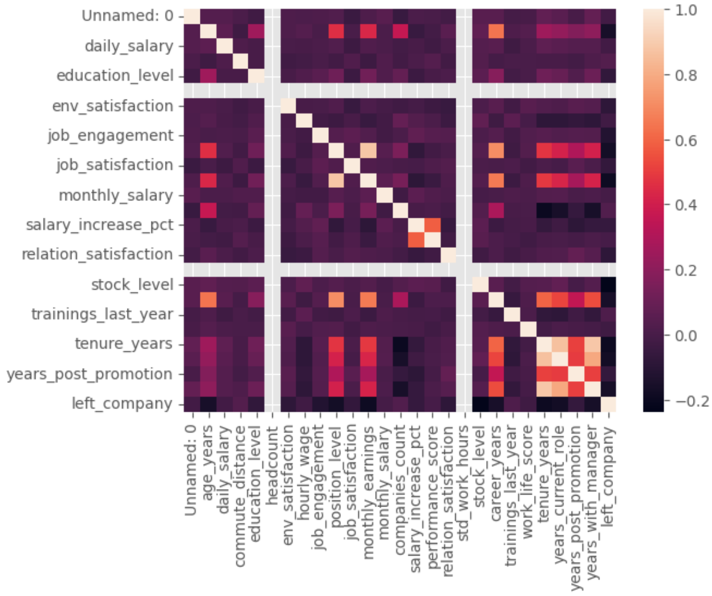

Figure 3: Heatmap of Correlation Matrix of the Features

4 Data Transformation

We will encode the categorical features in our dataset using *One-Hot Encoding*.

5 Handling of the Data Imbalance

We have used two different techniques together, in order to address the issue of **Data Imbalance**:

- Oversampling of the majority class via **SMOTE**. This has significantly improved the training & validation f1-scores by tens of percentages.
- Class weighting in any model in our analysis, that supports class weighting. This has improved the training & validation f1-scores by only a few percentages.

6 Model Training & Evaluation

6.1 Hyperparameter Tuning:

After Hyperparameter Tuning, we get to the following result:

Table 1: Hyperparameter Tuning Results

Model	**Val F1**	**Train F1**	**Time (s)**	**Best Parameters**
LR	88.0 ± 13.4	91.8 ± 2.6	6.20	clf__C: 0.01
SVM	89.9 ± 16.2	96.6 ± 1.5	12.36	clf__C: 1, clf__decision_function_shape: ovo, clf__kernel: rbf
LDA	87.6 ± 15.1	91.7 ± 2.5	2.89	clf__shrinkage: 0.3, clf__solver: lsqr
RF	92.7 ± 7.7	99.6 ± 0.2	131.33	bootstrap: True, max_depth: 15, max_features: log2, min_samples_leaf: 2, min_samples_split: 4, n_estimators: 100, n_jobs: -1, oob_score: True
AdaBoost	88.2 ± 9.7	91.8 ± 2.3	38.93	learning_rate: 0.5, n_estimators: 200
XGBoost	91.5 ± 11.2	100.0 ± 0.0	20.97	learning_rate: 0.1, max_depth: 5, n_estimators: 150, subsample: 0.8
LightGBM	92.3 ± 10.2	100.0 ± 0.0	159.86	colsample_bytree: 0.8, lambda_l2: 0.01, learning_rate: 0.1, max_depth: 10, min_child_samples: 5, n_estimators: 100, subsample: 0.8
CatBoost	92.1 ± 11.7	100.0 ± 0.0	204.58	depth: 6, iterations: 500, learning_rate: 0.1

6.2 Evaluation:

6.2.1 Accuracy & F1 Score:

Table 2: Accracy & F1 Scores for Training and Validation

Model	Dataset	F1 score Class 0	F1 score Class 1	Accuracy
SVM	Training	0.97	0.96	0.96
	Validation	0.93	0.20	0.88
LR	Training	0.92	0.91	0.92
	Validation	0.93	0.46	0.88
CatBoost	Training	1.00	1.00	1.00
	Validation	0.93	0.26	0.87
RF	Training	1.00	1.00	1.00
	Validation	0.93	0.35	0.88
XGBoost	Training	1.00	1.00	1.00
	Validation	0.93	0.28	0.87
LightGBM	Training	1.00	1.00	1.00
	Validation	0.92	0.24	0.86
LDA	Training	0.92	0.92	0.92
	Validation	0.93	0.39	0.87

6.2.2 ROC & AUC Scores:

Table 3: ROC AUC Scores Comparison

Model	Train AUC	Validation AUC
AdaBoost	0.9709	0.8409
CatBoost	1.0000	0.8332
LDA	0.9700	0.8097
LR	0.9706	0.8116
LightGBM	1.0000	0.7491
RF	1.0000	0.8265
SVM	0.9971	0.7543
XGBoost	1.0000	0.8172

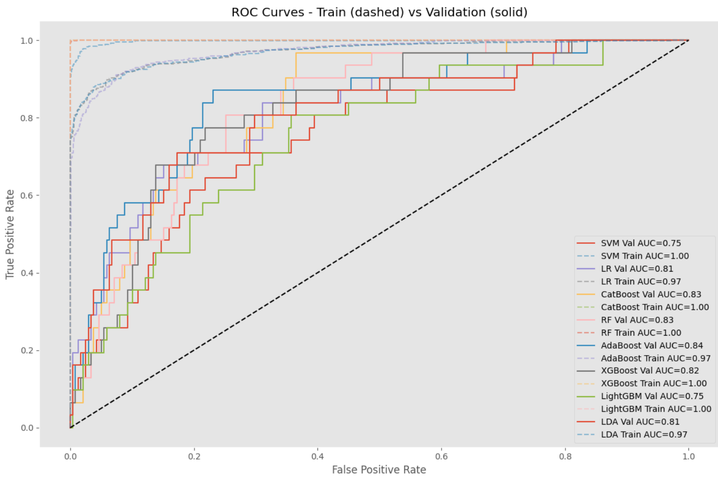

Figure 4: ROC & AUC scores for each of the models

6.3 Conclusion:

As we saw in **Evaluation** section, **Logistic Regression**, **Linear Discriminant Analysis** and **Random Forests** obtain a significantly higher f1-score on the minority class, compared to others, thus these are the best models for this task.