Payroll Predictive Model Discussion

1. Data exploration summary and important findings.

	Obs.	Missing Values	Wrongly labeled	Payroll rate	credit>0 labeled payroll	Unique Descrip. (total)	Unique Descrip. (payroll)	Unique Descrip. (non-payroll)	Rate of credit >10,000
Train	356,876	0	76	13.16%	59%	5,627	5,607	20	0.44%
Test	75,950	0	171	16.7%	65.7%	1,710	1,690	20	0.4%

1.1 No payroll has a transaction number in the description.
1.2 All non-payrolls have a transaction number in the description.
1.3 The transaction numbers in the tail of descriptions are almost unique and of no value.
1.4 After removing the transaction numbers from descriptions of non-payrolls, all non-payrolls are described with only 20 unique descriptions.
1.5 Both debit and credit have a 'power law'-like distribution with hundreds of outliers (assuming 10,000 as the threshold).
1.6 There are three main strong correlations between input variables and the target as follows. These are suspected to be data leakages.

- having/not having the transaction number defines the target perfectly
- credit values are highly correlated to target and can define it even without descriptions
- description is also highly correlated to target and can define it even without credit because non-payrolls (87% of data) are defined by only 20 descriptions

2. Data Cleaning

2.1 All records with coincident debit > 0 and credit > 0 (wrongly labeled) are removed (e.g. 76 in train data).
2.2 There are no missing values in train and test data.

3. Feature engineering

3.1 Four new variables were derived from Date. Day of month, day of week, week of month and month of year. These introduce stronger features in predicting payrolls.
3.2 Debit and credit were log scaled to alleviate the outlier effect.
3.3 Transaction numbers trimmed from the tail of descriptions.
3.4 To highlight nuance differences among descriptions a sentiment analysis was applied to convert them to embeddings (384 new features).

3.5 Feature selection applied to embedding using PCA to pick the top 50 PCAs (this number was chosen after analyzing the cumulative variance content of PCAs).

4. Model development and validation

4.1 XGBoost classifier that is a very efficient gradient boosting algorithm was used. It is not sensitive to the scale of features and is robust to missing values. Moreover, it is efficient in capturing non-linear relationships and interactions.
4.2 As the train data has more than 356k observations and target is not sparse, it was split into train (80%) and validation (20%). The test data (given as a separate file for the assignment) was used as the holdout data for testing and benchmarking.
4.3 Using grid search and cross validation the model's hyperparameters were tuned.
4.4. Overall and class accuracies, precision, F1 score and ROC-AUC were used as metrics with the focus to optimize ROC-AUC (because of class imbalance).
4.5 As the model showed extraordinary performance (suspicious to data leakage as described in 1.6 above), experiments were repeated with different combinations of variables (five combinations of features in total) as follows to verify where is the leakage from.

- 4.5.1 All features + description embeddings (PCAs)
- 4.5.2 All features, no description embeddings (PCAs)
- 4.5.3 All features, no description and no description PCAs
- 4.5.4 All features + description PCAs, no credit and debit
- 4.5.5 All features + description, no credit and debit or description PCAs

Note: where credit was removed, debit was removed as well because they have a mirror effect.

4.6 All above combinations yielded unexpectedly high performance on training and validation data with the 4.5.1 topping all (perfect class accuracies = 1)
4.7 Measuring the importance of variables confirmed that credit and description are of too strong importance in our model.

5. Model Testing

5.1 Same data exploration, cleaning and feature engineering/selection were done on the holdout data (test data).
5.2 The test data held untouched during the development (training and validation) were predicted using the model.
5.3 The best models obtained through 4.5.1 and 4.5.2 were employed on the test data.
5.4 Evaluation of the model proved that its performance on the unseen data is comparable to training and validation stages. The model's class 1 (payroll) accuracy is few percent lower though, its ROC-AUC is almost 1.

6. Benchmarking

6.1 A Logistic Regression Model as a strong baseline for interpretability and performance comparison was developed for benchmarking.
6.2 For the sake of simplicity and regarding the fact that the current datasets yield near perfect prediction accuracy, the cleaned and features engineered train and test data were used without sentiment analysis on the description variable.

7. Conclusion

7.1 Multicollinearity
The extremely strong predictive power of credit and description (that is unusual) overshadowed the full benefits of embedding description and date-driven features.
The possible reason can be that these datasets are made up (simulated) data. And, in the simulation the effect of multicollinearity is not avoided. The variables that strongly show this effect are ‘credit’, ‘debit’, ‘description’ and ‘is_payroll’.

7.2 Work on real data
If more time and resources are given, real data would be used with this model to measure its performance on real transactions. The applied sentiment analysis (embeddings) and the feature selection method (here PCA was adapted) would be further investigated to exploit the description variable. Although description embeddings pushed the model’s performance a bit, they were overshadowed by the unusual predictive power of the original variables.