FEATURE ENGINEERING REPORT

Victor Vu Duy Phuoc

Executive Summary

This report layout the pipeline of the feature engineering process to predict the whether an employee stay or leave the company. In section 1 and 2, we discuss the critical steps that were taken to prepare the base model:

- Converting boolean variables
- Performing integer encoding and one-hot encoding categorical varables
- Check skewness and fix skewness if appropriate
- Perform correlation analysis
- Perform feature reduction using Recursive Feature Elimination

Applying Logistic Regression model, we yield an accuracy score of 0.79 on base model. Section 3 focuses on the feature engineering process, where we construct 24 new features and improve the accuracy score to 0.98. We acknowledged the high risk of overfitting in our model and have applied PCA dimensionality reduction yielding an accuracy score of 0.97 with only 25 dimensions.

I. Exploratory Data Analysis

1.1 Data Loading

Data was loaded directly from computer path using the read_csv function inside Panda library.

1.2 Data Understanding

The dataset contains contains 14999 entries with no NA. There are 10 columns holding float, interger and object value type. Catergorical variables of "left", "work_accident", "promotion_last_5years" were encoded as numeric. Salary is populated binned as categorical of "low – medium – high". It is important to note that only 24% of the data represents employees who left, and the imbalance dataset might have an impact on the accuracy of the predictive model. However, solution to this issue will bot be discussed in this report, the focus of this project is to build robust model from feature engineering.

II. Building Base Line Model

In order to build the Base Model, the following data transformation were being performed:

- Convert Promotion_last_5years & Work_accident, which only hold boolean values of 0 and 1, into binary variables.
- Perform integer encoding on categorical variable "salaries" into numerical values of 0, 1 and 2. This is build up on the fact that values have natural ordered relationship where high salary is higher than medium salary, which is higher than low. Another way to deal with this variable is to apply one-hot encoding and create 3 new dummy variables, however this model generate lower accuracy score, which might be because some information about the ordinal relationship was lost in this transformation.
- Values in categorical variable "sales" has no ordinal relationship, thus it is appropriate to One Hot Encoding of "sales" variable. Each department is now a dummy variable.
- Check the skewness level of integer and float variables as Normality is a key assumption in this statistical model. Then apply boxcox method to if skewness is above 2. The values for asymmetry and kurtosis between -2 and +2 are considered acceptable in order to prove normal univariate distribution (George & Mallery, 2010)[1]. In this case, the highest skewness level is 1.85, so no skewness transformation was performed.

[1] George, D., & Mallery, M. (2010). SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a ed.) Boston: Pearson.

- Correlation analysis was performed to ensure the model does not included highly correlated variables. In this case, there is no highly correlated variables.
- To avoid overfitting, perform features reduction by using Recursive Feature Elimination and remove insignificant features with p-value more than 0.05. Five predictors were removed. The model is now left with 13 predictors.

Since the amount of data entry is not large, we will applies train test split of 80/20. Using a low split of data for training generates a lower accuracy score. Logistic Regression is chosen as we are trying to predict binary variable "left". Specifically, to further reduce over overfiting, we applies Ridge Regression or L2 Regularization as penalty of 0.5, which puts penalty on high coefficient but not eliminate coefficients as L1 does. The metric that we used to evaluate the model is accuracy, which is the number of correct predictions made as a ratio of all predictions made. Applying this model on the test set, we received an **accuracy score of 0.79**. Cross Validation with k-folds of 10 generates accuracy scores ranging from 0.77 to 0.81, thus validating the normal behaviour of this model.

III. Feature Engineering

Simple data analysis was perform to understand the relationship between each numerical predictor and the target variable "left". The observations from this analysis built the core of feature engineering.

3.1 Binning & One Hot Enconding "Satisfaction Level"

This is among the most important factor determining whether the employees stay or leave. From initial data exploration, the following trend was spotted out:

- Employees with satisfaction below 0.12 have left.
- Employees with satisfaction from 0.36 to 0.46 have left.
- Employees with satisfaction from 0.92 and above all remain.

First attempt of binning was done by constructing 10 equally divided percentile of data distribution from 0 to 1. These bin categories would then be converted into dummy variables. This generates an improvement in Accuracy to 0.82, however this new features still capture the non-linearity behaviour of this predictor.

Thus, to remove non-linearity, these ranges are binned into 5 bin range that behave similary. [0 : 0.12] - [0.12 : 0.36] - [0.36 : 0.46] - [0.46 : 0.92] - [0.92 : 1]

→ **Accuracy score resulted: 0.88 (Accepted)**

3.2 Binning & One Hot Enconding “Last Evaluation”

From initial data exploration, the following trend was spotted out:

- Employees with evaluation below 0.48 have left.
- Employees with evaluation from 0.66 to 0.89 have high likelihood of leaving.
- All employees with evaluation above 0.9 remains.

Similar to the above transformation, binning by distribution percentile yields very minor improvement due to non-linearity. Thus, 4 bin levels that behave fairly similar were constructed. [0 : 0.48] - [0.48 : 0.66] - [0.66 : 0.9] - [0.9 : 1]

→ **Accuracy score resulted: 0.90 (Accepted)**

3.3 One Hot Encoding “number_project” and “time_spend_company”

One hot encode “number_project” and “time_spend_company” into dummy variables. These transformations help remove the nonlinearity of these two predictors.

→**Accuracy score resulted: 0.945 (Accepted)**

Another attempt to create new features from these two predictors is to bin similarly behaving values. However, this model yields lower accuracy score, which might due to the fact that the bin categories contain different behaviours and non-linearity has not been fully resolved.

→**Accuracy score resulted: 0.94 (Rejected)**

3.4 Binning & One Hot Enconding “Monthly_hours”

This variable show the number of hours the employee work per month. The following observation were made from analysing working hours with target variable.

- Employees with less than 132 hours of work have left
- Employees work from 166 hours to 259 hours per month is likely to leave
- All the one work above 260 hours remain.

Similar to the first two feature engineers, binning by distribution percentile yields very minor improvement due to non-linearity. Thus, 7 bin levels that behave fairly similar were constructed. [0 : 132] - [132 : 165] - [165 : 178] - [178 : 179] - [179 : 259] - [259 : 287].

→ **Accuracy score resulted: 0.947 (Accepted)**

It is arguable that this new feature only improve the model slightly however adding complexity and variance to the model. As we apply polynomial features to this model, it yields a higher accuracy score of 0.98. The improvement could be because polynomial features help further explain all the non-linearity of the model. Whereas if we run polynomial features on the model without this feature engineering, we get an average score of 0.95, which might underfit and include higher bias. In this case, we will accept this new feature taking into account of the risk of overfitting.

3.5 Outliers Detection and Cross Validation on New Model

Leave-One-Out methodology with Ordinary Least Squares was employed to find which points have a significant effect on our model fit and remove outliers. However, no significant outliers were detected.

Applying the new model on the test set, we received an **accuracy score of 0.95**. Cross Validation with k-folds of 10 generates accuracy scores ranging from 0.93 to 0.96, thus validating the normal behaviour of this model.

3.6 Apply Polynomial Features

Applies polynomial features to further resolve the nonlinearity among all variables generates an **Accuracy score of 0.98** on average of 10 k-folds. Then, we obtain scores and estimators from different splits and use the best one. Using a polynomial function of your predictive features allows the model to fit a non-linear, more expressive decision boundary. The better performance suggests that the data are better separated this third-order polynomial boundary. However, we are risking overfitting our model.

→ **Accuracy score resulted: 0.98 (Accepted)**

3.7 Dimensionality Reduction using Principal Component Analysis

The model now contains 37 predictors which might express redundancy properties. Thus to reduce the dimension or complexity of the model, we will try PCA which uses the variance of each feature to find new features in order to maximize its separability. There is no improvement to the Accuracy score, however we yields an **accuracy score of 0.97** with only 25 features. This is 12 dimensions less than our previous model. Thus, this could be an appropriate solution to be considered for reducing the complexity of our model.

Conclusion

From the base model of 13 dimensions yielding an accuracy score of 0.79, we have raised the accuracy score to 0.98 by constructing 24 new features. Since the data of the base model is non-linear, Logistic Regression was underperformed. Thus, the main task of feature engineering is creating predictors that hold linear relationship with the dependent variable “left”. This was achieved by grouping values with similar behaviour.

While our best model yields an accuracy score of 0.98, we need to take into account of the risk of overfitting the model with polynomial features. To reduce the complexity of the model, we have tried dimensionality reduction using PCA and got an accuracy score of 0.97 with 25 dimensions.