Approach – Classification of Business and Personal Expenses

Business and personal expenses get grouped together due to improper entries/ reporting of expenses. This document outlines an approach which combines series of rules and features, external to the expense data to arrive at a conclusion, whether an expense is a business or personal. The rules and features considered in addition to features extracted from expense descriptions are given below:

- Role of an employee and corresponding expense eligibility
- IRS guidelines on acceptable business expenses (https://www.irs.gov/publications/p535#en_US_2017_publink1000208614)
- IRS guidelines on acceptable (meals, entertainment and travel) Business expenses https://www.irs.gov/pub/irs-pdf/p463.pdf
- Holiday lists and expected expense patterns during holidays

1. Introduction

This is an approach paper outlining an algorithm to help discriminate between the business and the personal expenses. Currently the expenses are split into micro categories such as

- Travel
- Meals and Entertainment
- Computer – hardware
- Computer – software
- Office Supplies etc.

The algorithm proposed in this paper incorporates, certain other information outside of the expenses description, to classify whether an expense is “Business” or “Personal”.

In this paper, the categories will be limited to the above stated.

2. Extracting Value from Existing Data

Before enriching the data, maximum information should be extracted it. Some of the categories are very explicit and have reasonably smaller margin for error.

Example: Office Supplies

Expense items under this category can be classified as “Business”.

2.1 Isolating Obvious Business Expenses

Computer – Software, based on the historical data, software for business can be added to a watch list or a look up table. When an expense item is logged and a match is found within the watch list, then that expense item can be logged as either “Business” or “Personal”

Example: Microsoft Office, Adobe Acrobat DC, Drobox etc.

This approach aids in filtering out the expense items which can be easily classified into either of the desired buckets. Given that data under two sub-categories are classified, the reduced dataset is taken up for further processing.

Similarly, any expense item with the mention of family will get classified under “Personal”.

2.2 Data Cleanup

Before the data gets passed down the pipeline, data quality should be ensured.

Example: Duplicate records for “Jonathan Ive” in employees.csv file

Duplicate employees and other duplications should be found and dealt with appropriately. In case of mentioned example Jonathan has 2 employee Ids 2 and 5, easier approach is to retain records having employee Id as 5 but

change the ID to 2. In summary replace duplicate entries by first identified.

3. Role – Based Expenses Eligibility

Next step is to assign role-based expense eligibility matrix. This matrix will operate as a discriminating feature. An illustrative example is given below:

	Computer	Iphone	Team Lunch	Dinner	Taxi	flight
CEO	X	X	X	X	X	X
Engg	X					
Sales	X	X	X	X	X	X
IT & Admin	X					

From the table, it becomes easy to identify, who is eligible for what kind of expenses. This matrix will help in classifying the expenses better.

Additionally, how many expenses of a given category is allowed for an employee per year, should also be decided. It will help avoid redundant expenses of same nature.

Example: An employee buying more than one iPhone under computer-hardware category, whereas the eligibility is to buy a phone once every 2 years.

3.1 Role Based Expense Ceiling

For each role a threshold need to be set for each category of expense. A CEO might entertain the team with team lunch and might have extra privilege and expense budget as against an engineer.

Example: In the given employees.csv, it can be observed that the sales team have more expense under the “Meals and Entertainment” category. It is applicable to the role, as they might meet either and existing or a prospective client over a lunch/ dinner or a coffee. Whereas an engineer, having a coffee (if not in company policy) in Starbucks will be “Personal”.

Similarly, we can see 2 sales guys have expensed Macbook air, one costs 2000$ and the other 4000$, assigning a ceiling on the eligible amount that can be claimed, will help categorize better.

4. IRS Guidelines

IRS, have provided clearly defined guidelines for what can be considered as business expense and what cannot be. This data is available in the IRS website.

Example: https://www.irs.gov/pub/irs-pdf/p463.pdf , it can be observed that a person residing in the same city as his/her place of work, cannot claim Taxi expenses under “Travel”. It can be observed for employee number 7 in the dataset.

Hence, the IRS guideline pdfs and the website can be used as a data source. The Pdfs should be parsed and the website scraped for relevant data. The data thus obtained should be cleaned and processed. Natural language processing (NLP) algorithms can be used to generate features from this text and can be added as IRS feature to each expense item.

5. Holiday Lists and Expense Patterns

It is essential to look at the transaction dates to identify expenses filed during known holidays. It will also help in understanding the seasonality of expenses filed during holidays.

Example: Employee Id 4, with sales has maximum number of “Meals and Entertainment” items. It can be observed that there is an expense logged on 31-Dec for 1000$, which is way above his regular expenditure pattern, observed through out the year. It is a possibility that this a personal expense.

It also helps setup a threshold around the observed spending of individuals under a specific category. Any significant deviations from the established pattern can help in identification of one class from another.

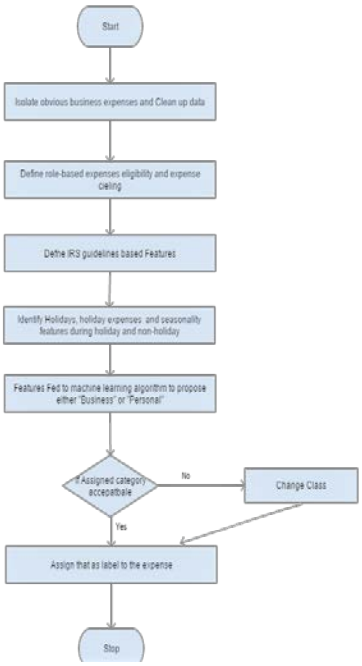
6. Machine Learning

Once the features are generated from steps 3 till 5, they can be used to develop a recommender system.

The system can recommend the category to be either "Business" or "Personal", if the end-user accepts the recommendation, it is a true identification otherwise a false one. This helps in generation of labelled dataset and also provides feedback for tuning.

This labelled dataset can be used as input for a classification algorithm.

7. Algorithm Flowchart



8. Conclusion

This approach is a combination of experience, domain expertise and machine learning. As in any machine learning project, the success is dictated by the diligence followed in feature engineering and model tuning.

Here Section 3 – 6 can be run in parallel, as they are independent of each other. Hence, a distributed feature engineering will aid in accelerate the model building process.

Finally, the model needs to get feedback and retune itself. It is advisable to start with human in the loop feedback system and move on to automated feedback and reinforcement.

9. References

- IRS guidelines on acceptable business expenses (https://www.irs.gov/publications/p535#en_US_2017_publink1000208614)
- IRS guidelines on acceptable (meals, entertainment and travel) Business expenses https://www.irs.gov/pub/irs-pdf/p463.pdf
- https://bentoforbusiness.com/personal-vs-business-expense-deductions/
- https://budgeting.thenest.com/tax-laws-computer-expenses-deductions-24866.html
- https://www.thebalancesmb.com/caution-with-business-tax-deductions-3866098