Spotle.ai Study Material
Spotle.ai/Learn

Corporate Expense Auditor

An Application Area Of Machine Learning

Overview:

Travel and entertainment expenses are one of the most frequent expenses in a company. And these expense categories can be mis-used or manipulated, if not thoroughly audited. The question is can we not do a proper checking to identify potential errors or abnormalities or fraud? Surely this can be by applying proper techniques to stop misuse of travel & entertainment expense policy and shielding the expenses from potential fraud.

This project aims at developing a statistical algorithm which highlights the potential error or misuse of policy in travel & entertainment expense data of any multinational company.

The transaction data are logged through Expense Reporting systems by the employees. These data will be analyzed using statistical model to identify violations of the corporate expense policy or fraud and abnormalities. Henceforth, it would help the Prepay Audit team to perform efficient auditing and take appropriate corrective actions.

The company may use any Enterprise Service Automation for travel

and expense management which automates travel & expense process and establishes policy-driven controls for expense reimbursement. It also provides web portal where employee can access services for travel and expense settlements.

Travel & Expense Process

Booking air-tickets/ hotels/ cabs/ rail-tickets etc on company work purpose. The travel expenses incurred for employee's business trip are managed by an integrated management system. A business trip can be of short or long duration depending upon the project requirement. Based on the travel type employee can place travel request for flight tickets, hotels and cabs etc. Within 30 days of completion of the business trip, employee has to submit the expense report with necessary data or receipts for reimbursement and settlement of cash advance.

Standard Process of any Business Trip

1. Once Visa is approved for travelling, employee submits request for ticket booking and accommodation.
2. After PM's or BU Head's approval, the travel request is routed to Travel Desk. They arrange for air tickets, hotel accommodation and local taxi for commute in accordance to employee's eligibility.
3. Employee applies for insurance.
4. Global Immigration Team arranges FOREX which employee can use for immediate expenses after travel.
5. Employee pays for hotel and taxi charges using card or cash.
6. For all card transactions, employee has to submit expense report through internal portal.
7. Employee has to submit claim report for reimbursement of business expenses paid by cash.

Expense Reporting System

Expense reporting system is divided into four groups based on

geography. India, US, UK and APAC. US region includes North America, Canada, Brazil, Costa Rica and Argentina. UK geography covers United Kingdom and Europe. APAC region includes Asia, China and Middle East countries.
About 30,000 expense reports are generated each month. In terms of Report Volume share, US holds 45% and India has 35% of the total expense report volume. Each geography has customized expense report template as listed in the below diagram. Each template has many expense type listed in it.

Each template is tied to an Expense or Financial Policy. For Instance US – Initial stay template is associated with Initial Stay policy. All expense claims under it should adhere to that policy.
US Geography:

- Initial Stay Template => I/S policy
- Domestic Travel Template => Business policy
- Relocation Template => Relocation Policy

Cash expense can be reimbursed using claim process.

Status of the Expense Report

Status Code	Status Description	With
NA	Unassigned	Employee
PND	Pending	Employee
SUB	Submitted	Audit Team
PAR	Pending for Approval	PM
STGD	Ready for Payment	Bank Processing

Objective

1. To build a system called a 'Corporate Expense Auditor' based on historical/past expense data.

2. Input to the system would be expense data.

3. System will add an additional column named 'Audit'.

4. If the value of Audit = 'NO', then auditor may approve the report.

5. If the value of Audit = 'YES', then auditor should audit the report take appropriate corrective actions.

Methodology

Ensemble methodology would be adopted which is an ensemble of business rule based classifiers and machine learning algorithm based classifier.

1. Develop the business rule based classifier.
2. Develop machine learning based classifier.

Classification based on Business Rule

Business rule based classifier will be developed based on the direct help from pre-pay auditors.

- For example: for NA_BUSINESS template for expense type cu0000 if amount is greater than $200 and merchant type is

unknown then we will raise the flag like Audit = YES - otherwise Audit = NO.

- For the expense type where the business rule is not properly defined, either a k-sigma rule or empirical limit(s) will be applied.

- The value of k or the empirical limit(s) will be estimated based on the historical data.

- For each expense type, merchant description data dictionary will be created. It will be an .excel or .csv data file with each expense type as column and list of valid merchant description. The merchant data will be populated using historical expense report data.

Classification based on Machine Learning Algorithm

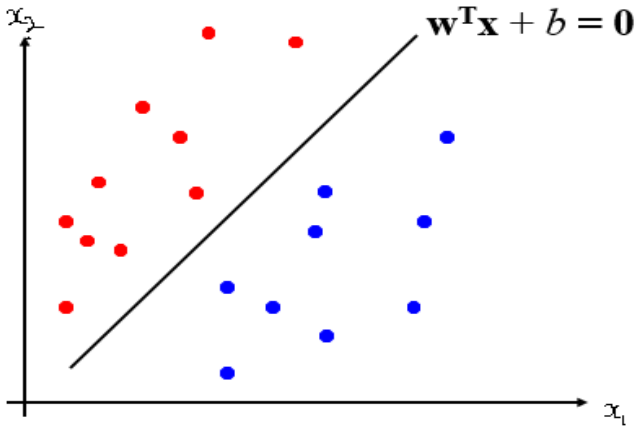
- Historical data will be randomly split into train and test data.
- Following machine learning algorithms will be implemented
  (i) Support Vector Machine (SVM),

(ii) Decision Tree

(iii) Logistic Regression

- The efficacy of these algorithms will be tested on the test dataset
- The process will be run on several datasets
- The best algorithm will be selected and an audit score will be generated in a scale of 0 to 100. Following rules would be followed:
  - If score > 50 then Audit = YES
  - If score <= 50 then Audit = NO

i) Support Vector Machine (SVM)



In machine learning, support vector machines (SVMs) are supervised learning models that analyze data and recognize patterns, used for classification analysis. For example, in the attached figure $x_1 = \text{Age}$ and $x_2 = \text{Income}$ are two features to classify the good and bad loans. Suppose red-dots indicates bad loan and blue dots indicates good loan.

Bank wants to classify the loans into two categories - good and bad loan. Based on two features x_1 and x_2 a line $w^T x + b = 0 <=> w_1 x_1 + w_2 x_2 + b = 0$ is built. The line is known as classifier and it classifies the good and bad loans.

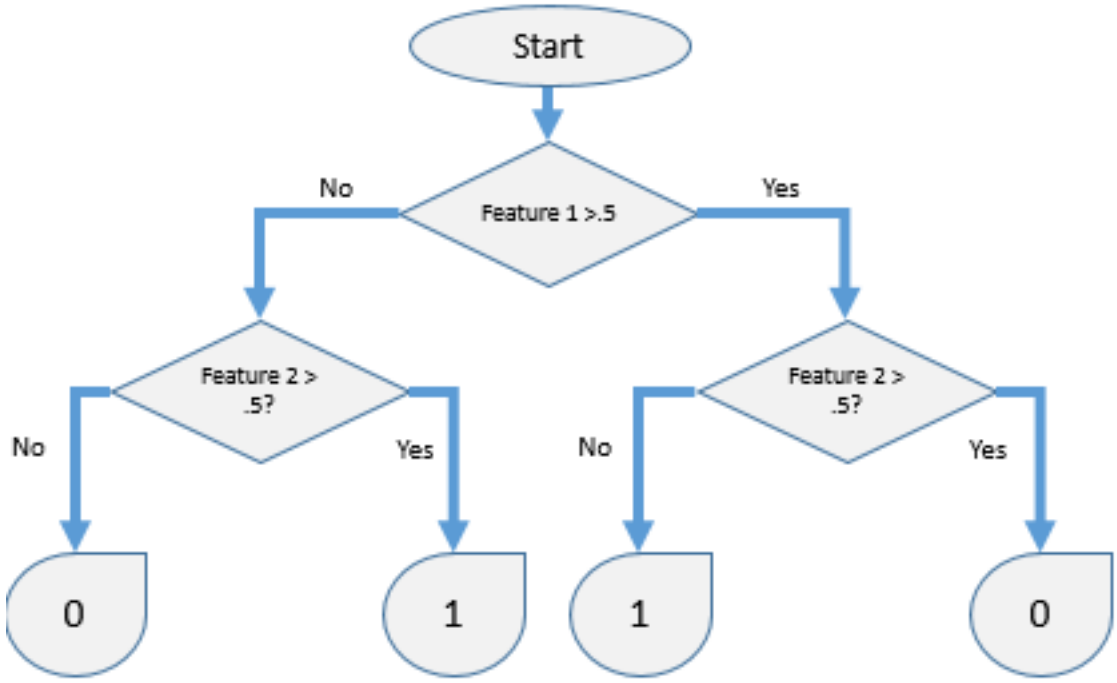
ii) Decision Tree

A decision tree is a popular data mining decision making tool that uses a tree-like graph. It is one way to display an algorithm.

The objective is to create an algorithm that predicts the value of a target variable based on several predictor variables.

Each interior node corresponds to one of the predictor

variables. There are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.



iii) Logistic Regression

Logistic regression is probabilistic classification model. Suppose

$$P(C_1 = \text{good loan} \mid x) = \frac{\exp\{x^T\beta\}}{1 + \exp\{x^T\beta\}}$$

For a given set of new feature variables, if

$$P(C_1 = \text{good loan} \mid x) > 0.5$$

then the new case will be considered as good loan. Otherwise it will be considered as bad loan.

iv) Profile Attributes based Scoring

This idea is more close to the idea of clustering. Based on different profiles of the report, such as Expense_Template, Grade_Bucket, group-id, and Project_ID, transaction pattern will be modeled using probability models and distance score for given profile will be created for each transaction. High score indicates that the transaction is out of the cluster and it will be flagged as exception.