tide

Receipt Matching Data Science Challenge

1. Description

Our customers want the ability to automatically match receipt images to the associated transaction within the tide app. For this we have used an external supplier to extract data from the receipt images, (for example, the date of the transaction on the receipt, the name of the merchant, the transaction amount). Each of the data items returned from the external supplier was compared against the same data from the tide transaction, and a 'transaction-receipt' matching vector was created. The elements of the matching vector will be produced in different ways depending on the underlying data types being matched (for example for strings it may be a fuzzy matching score, for transaction amounts it may be an absolute difference of the amounts on the receipt and the transaction, for dates/times it may be a time delta, some may be discretized measures of confidence).

Since the data extraction from the receipt image is not always perfect (for example the incorrect string is extracted for the merchant name) we want to build a model to learn which matching features are the most successful. The ultimate goal is to match a single receipt to the correct transaction given a number of possible transactions however, given real world considerations, we want to sort the possible transactions for a given receipt in order of likelihood of being the correct transaction. So in the app when the customer takes a picture of a receipt, the app provides a list of transactions likely to match the receipt, with the one we think is correct at the top of the list. 'Success' in this context means that the correct transaction for the given receipt is at the top of the list, (note, if the correct matching is not in the data for a given receipt 'success' is not possible).

2. Deliverables

Model

Your code should produce a trained model which can be used to order a set of transactions by likelihood of matching a receipt image. This model should be able to take a number of 'transaction-receipt' matches and order them with the most likely to be correct at the top of the list.

Code

You need to deliver the code used for this work as a git repo. The code should be written in python 3, you may use any additional packages available in the python package index.

Report

You should include a short report explaining the approach you took and why, the results of the model and some discussion of what they mean and any recommendations you would give to the team to improve the results (assume the reader is a software engineer with a computer science background but with no specific expertise in data science/machine learning). The report may be in a separate text file or within a jupyter notebook. It should explicitly state the success of the model as experience by the end user.

3. Data

The data consists of a csv file where *'member_id'* is the tide customer identifier, *'receipt_id'* is the unique identifier for a receipt image, *'matched_transaction_id'* is a unique identifier for the transaction that we know is the correct match for the receipt_id, *'feature_transaction_id'* is the unique_identifier for the transaction which was combined with the receipt_id to produce the matching vector. The rest of the columns are the elements of the matching vector for the given receipt_id and matched_transaction_id. Note some filtering was performed in an attempt to reduce the number of receipt-transaction matching vectors therefore not every receipt was matched with every transaction for the member.

For example

receipt_id	member_id	matched_transaction_id	feature_transaction_id	Matching columns...
1234	abcd123	xyz	qwe	
1234	abcd123	xyz	xyz	
1234	abcd123	xyz	cvb	
7890	abcd123	cvb	cvb	
7890	abcd123	cvb	xyz	
7890	abcd123	cvb	egh	
7890	abcd123	cvb	iop	
5678	bcde234	wsd	qaz	
5678	bcde234	wsd	ygv	

This data shows us that for customer abcd123, receipt_id 1234 had 3 possible matchings which passed the filtering, the correct transaction which maps to this receipt is transaction xyz, which passed the filtering and is in the data set (row 2) and also two other transactions passed the filtering for this receipt_id (transactions qwe and cvb) but these are incorrect matchings.

For member abcd123, receipt_id 7890 had 4 possible matchings which passed the filtering, one of which was the correct match (row 4) the rest are incorrect matchings.

For member_id bcde234 there is only one receipt_id 5678 for which two transactions passed the filtering none of which were correct matchings.

Good luck!