ワールドジャパン学習用データの整理作業　引継ぎ資料

1.　学習によって導く内容

各顧客の基本データ・アンケート回答結果と来店日リストを用いて契約期間を満了した顧客の要因を学習によって導く。先方の希望の学習に用いるアンケート項目は来店年月日、生年月日、年齢、職業、未婚/既婚、知った理由、家族構成、DM 希望有無、美容に使う金額、他サロンに行ったことあるか、性格の 11 項目だが、欠損値が多く、有効データが非常に少ない。そのため、とりあえずとして比較的にデータの欠損が少ない来店年月日、生年月日、年齢、職業、未婚/既婚の 5 つの項目に絞ってデータ整理を行い、有効データを判定する。データ整理によって得られたデータを分析が可能な形式に加工し、主成分分析を行うことで、契約期間満了の要因を導く。

現時点で、絞った項目でのデータ整理、有効データの集計まで終了している。今後は、職業などの非数値を数値化するなどの、学習データを学習に最適化するためのデータ加工作業と、実際にデータを用いた主成分分析を行う必要がある。

2.　学習用データの種類とデータ整理用プログラムの保管場所

＜学習用データの親フォルダ＞

S:¥個人作業用¥アルバイト¥ワールドジャパン¥学習用データ

＜各学習用データの格納先フォルダ＞

2 フォーマット、2 フォーマット 2、2 フォーマット(潮永さん)、2 フォーマット(渡部さん)、2 フォーマット(柏葉さん)、慶野さん、佐藤さん

＜各学習用データの種類＞

・【A】フォルダの中の xlsx ファイル

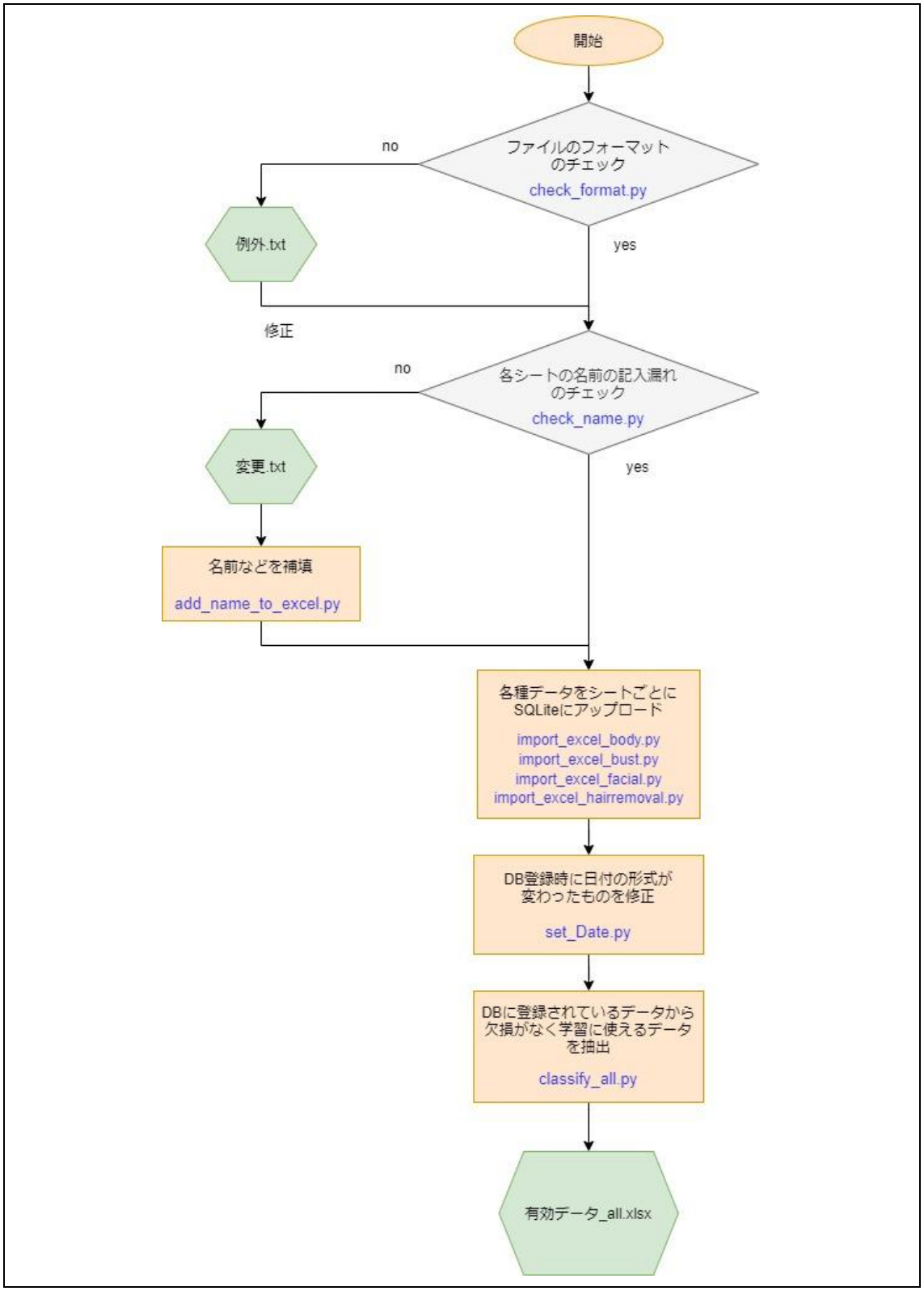
各顧客の来店時アンケートの回答内容。コースによってシートが分けられており、各シートに顧客の基本情報やコースごとのアンケートの回答が記載されている。このデータをシートごとにＤＢのテーブルに保存し、学習項目に欠損値のないデータのみ抽出して学習用データとする。

・【H】来店日リスト.xlsx
　各顧客の来店日のリスト。有効データとして抽出されたデータのうち、このリストに契約回数分来店日の記載があるかを確認することで契約期間の満了を判定する。

＜データ整理用プログラムの親フォルダ＞
　S:¥個人作業用¥那須¥ワールドジャパン¥now
　（S:¥個人作業用¥那須¥ワールドジャパン¥old には、現在のデータ整理形式に変更になる前に用いていたプログラムが保管されている。もしかしたら使えることがあるかも。。。）

3.　データ整理のフローと対応するプログラム

各プログラムの詳細な解説は各プログラム中のコメントを参照。



3.1. ファイルのフォーマットのチェック

使用プログラム：check_format.py

各ファイルのシートが稀に 1 行少ない・多いことがあるためそれをチェックする。フォーマットが異なるファイルはファイルパスが例外.txt に出力されるため、そのリストをもとに手動でフォーマットを整える。あまり多くはないため、現時点では自動で修正するスクリプトは作成していない。

3.2. 名前などの基本情報の記入漏れのチェック

使用プログラム：check_name.py

1 人が 2 つ以上のコースのアンケートに回答していた場合、あるシートには基本情報とアンケートの回答が記載されているが、別のシートにはアンケートの回答のみで基本情報が記載されていないことがある(例：ボディのシートにはすべて記載されているが、バストのシートにはアンケートの回答しか記載されていないなど)。後述する DB へのアップロードにおいて、各シートの名前の欄が空欄かどうかでそのシートがアップロード対象かを判定するため、補填しておく必要があり、情報補填が必要かのチェックを行う。

情報の補填が必要と判断されたファイルは変更.txt に出力され、次の処理にて補填が行われる。

3.3. 記入漏れの基本情報を補填

使用プログラム：add_name_to_excel.py

基本情報の記入漏れがあると判定されたファイルを変更.txt をもとに開き、別のシートから情報の補填を行う。

3.4. 各種データの DB へのアップロード

使用プログラム：import_excel_~~~.py

各シートに情報が記入されているかの判定を行い、記入されていた場合にはその情報を DB へアップロードする。シートごとにファイルを作成しているが、マージして 1 つのファイルにするといい感じかもしれない？(コードの行数が結構すごいことになる)

3.5. DB 登録時に日付の形式が変更されてしまったものを修正

使用プログラム：set_Date.py

Excel の日付形式データを DB へ pandas を用いてアップロードした際に、正しい日付形式でなく、誤った日付として保存されてしまうことがある。これは、Excel などのソフトウェアが日付を内部的に 1900/1/1 から経過した日数(シリアル値という)として保持していることが起因している。何の処理も行わずに Excel 内の日付をアップロードするとこの数値になってしまうため、openpyxl の関数によって数値を日付に変換してからアップロードする

が、この処理の際に稀に誤った日付に変換されてしまうことがある。DB 内の誤った日付を判定し正しい日付データで上書きする。全シート分の処理がブロック分けして記述されているため、各シートをコメントアウトすることで任意のシートのみの実行も可能。

3.6.　有効データを集計

使用プログラム：classify_all.py

学習に必要な項目がすべて欠損値でないデータを有効データとして抽出する。DB のテーブル別に処理を記述しており、この中のデータ抽出に用いているカラム名を変更することで学習に用いる項目の変更に伴うデータの抽出が可能。

4.　SQLite のテーブル構造

SQLite は“S:¥個人作業用¥那須¥ワールドジャパン¥now¥sqlite3“にあり、同フォルダ内の”salon_A.db”が学習用データの格納に用いている DB ファイルである。おそらく各 PC にインストールされている”DB Browser (SQLite)“を用いることでテーブル設定やデータ閲覧が可能である。

各シートのデータの格納にシートごとに 1~3 までのテーブルを用いている。各テーブルの役割を以下に示す。各テーブルのカラムは、実際の各シートの項目の順番に対応しているため、見比べることで大まかに内容を把握することができる(と思います)。

テーブル名	シート	説明
questionnaire_body	ボディ	“2 フォーマット”内のデータを格納
questionnaire_body2	ボディ	“2 フォーマット 2”内のデータを格納
questionnaire_body3	ボディ	上記 2 つ以外のディレクトリ内のデータを格納(追加分) ※
questionnaire_bust	バスト	“2 フォーマット”内のデータを格納
questionnaire_bust2	バスト	“2 フォーマット 2”内のデータを格納
questionnaire_bust3	バスト	上記 2 つ以外のディレクトリ内のデータを格納(追加分) ※
questionnaire_facial	フェイシャル	“2 フォーマット”内のデータを格納
questionnaire_facial2	フェイシャル	“2 フォーマット 2”内のデータを格納
questionnaire_facial3	フェイシャル	上記 2 つ以外のディレクトリ内のデータを格納(追加分) ※
questionnaire_hairremoval	脱毛	“2 フォーマット”内のデータを格納
questionnaire_hairremoval2	脱毛	“2 フォーマット 2”内のデータを格納
questionnaire_hairremoval3	脱毛	上記 2 つ以外のディレクトリ内のデータを格納(追加分) ※

※（後から追加されたデータである“2 フォーマット(潮永さん)”、“2 フォーマット(渡部さん)”、“2 フォーマット(柏葉さん)”、“慶野さん”、“佐藤さん”の 5 フォルダ内のデータを格納）

5. メインのデータ整理作業には用いないプログラム

- duplicate_rejection.py

  本当に稀にテーブルの UNIQUE 条件によって重複したレコードが登録されてしまうことがあるため、一回各テーブルのすべてのレコードをデータフレーム化し、重複を削除し、そのデータフレームを再アップロードする。

- trans_date_num.py

  このプログラムの引数としてシリアル値を指定することで、正しい日付を返す。ＤＢに日付がシリアル値でアップロードされてしまったものを手動で修正する際に使用。