**Implementing an Automation tool**

**to improve**

**the Quality of customer Data for PiLog groups.**

Submitted by:

Mohammed Mutahar Shaik **(**[**mmshaik1@ualr.edu**](mailto:mmshaik1@ualr.edu)**)**

Committee Chair & Faculty Advisor: Dr. John R Talburt **(**[**jrtalburt@ualr.edu**](mailto:jrtalburt@ualr.edu)**)**

Sponsoring Organization: PiLog Groups

Committee Member & Sponsoring Supervisor: Leon Claassens (**[leon.Claassens@pilog.co.za](mailto:leon.Claassens@pilog.co.za)**)

Table of Contents:

Executive Summary.………………………….…………………………………........3

Introduction………………………………………………………………………........5

Project Approach……………………………………………………………………....6

Project Process……………………………………………………………………........7

Data Quality Issues……………………………………………………………………8

Conclusions………………………………………………………………………........29

Acknowledgement……………………………………………………………………..30

References……………………………………………………………………………...30

1. **Exclusive Summary**

The purpose of this project is to implement a tool to improve the data quality of customer data belonging to PiLog Groups.

The project mainly focuses on three dimensions of data quality:

1. completeness,

2. validity,

3. Uniqueness.

The project involves generating synthetic data using a website called mockaroo <https://www.mockaroo.com/>  for testing the data based on the business requirements of the PiLog group before using in the Organizational Dataset of Vendor and customer datasets, because of ethical and privacy issues.

Further It goes through the process of the CRISP-DM (Cross Industry Standard Process for Data Mining.

Further we focused on Understanding the business problems and thereby selecting the Attributes that helps for improving the value to the organization.

and then going through the process of Exploratory Data Analysis using Pandas Profiling and python libraries to visualize and understand the data.

We classify the Customer data records based on the latest transaction history of Purchases in to two as:

1. Active records
2. Inactive records.

Active records is classified if the latest Transaction history falls within the range of 1 year, and inactive as above 1 year. Further we focus on these inactive records to attract the customers by analyzing the data and focusing on the Data Quality dimensions.

Thus, we enter into improving the Data Quality of completeness

and using machine learning libraries and entity resolution models to improve the data quality.

1. The raw data we considered based on the business requirements were having incomplete data in few attributes and especially those which effects the business, among them the attributes like Largest Bill amount, State, City, names and especially mail-id was the challenging task.

As now a days to get in touch with the customers and vendors the official mode of communication is mail-id.

To improve and fulfill these communication gaps of incompleteness we follow few rule-based approaches to reduce the incompleteness in the customer records.

Likewise, the next dimension we are focusing on is “Validity”.

In this Scenario our main aim is to improve the validity percentage in the customer records by focusing on the few attributes of Date of Birth, Social Security Number and Mobile numbers.

Every organization has their own definitions for validity but in general, data are considered as a valid if it conforms to the syntax based on format, type, range of the data.

This we focused on improving the Validity percentage of the data by following the rules and formats based on the US standards followed by PiLog Groups.

The Final dimension we try to improve is the Uniqueness of the data, by removing the duplication of records by linking the records and clustering the similar records and verifying weather the person is belongs to the same entity or not.

To improve this Uniqueness percentage, we try with three different methods.

1. Using the Record Linkage Python library and then calculating the Entity resolution metrics based on the Oyster (rule-based approach).
2. Using the probabilistic record linkage using an existing python library and then calculating the ER metrics using the SVM ML classifier
3. Using SPlink (Python package) for probabilistic record linkage (entity resolution) based on pairwise matching the records and deduplicate it without using the unique identifiers.

By comparing the ER metrics based on all the three methods we concluded the best way to improve the uniqueness.

Visualization was added up along with these calculations to have a better understanding when dealing with large business datasets.

The project will also create interactive dashboards to visualize the data and provide valuable insights. The deliverables of the project include detailed analysis of data, detailed descriptions of methods and approaches, output data in required format as requested for PiLog Groups.

**Introduction:**

Data quality is crucial for businesses as it affects customer experiences, operational efficiency, and regulatory compliance issues. Poor data quality can lead to inaccurate, incomplete, and unavailable data, which hinders business decisions and customer relationships.

These Poor data Quality leads to acute problems and effects the business. This reaction leads to a rush reactive approach to identify, evaluate, and purchase the technical solutions which may or may not solve the problems and in addition it leads to ignoring the root cause for poor data Quality.

Finally, organizations need to be able to effectively establish and communicate to the business client community the level of confidence they should have in their data. To do this, organizations need a way to formalize data quality expectations as a means of measuring the conformance of data to those expectations; they also need to be able to baseline the levels of data quality and provide a mechanism to identify leakages and analyze the root causes of data failures.

The capacity to quantify the costs associated with data problems in the day-to-day functioning of the organization is essential for building a business case for data quality improvement. The responsibilities of segmenting them across impact dimensions and identifying each impact inside lower levels of a hierarchical taxonomy make it easier to examine the detrimental financial effects directly related to "bad data."

It is possible to prioritize the correction of data issues by reviewing the scope of data failures considering their correspondingly unfavorable financial repercussions, which in turn depends on data quality tools and technology.

The difficulty in using the notion of "return on investment" to support the funding of an improvement project is the inability to track improvements achieved through the project over time to determine whether they are facilitating the promised positive benefits.

The project aims to improve data quality for customer data belonging to PiLog Groups by focusing on three dimensions of data quality -Completeness, Validity and Uniqueness.

**Project Approach:**

The project approach involves generating synthetic data using a mockaroo website and then later it is used for Organizational data.

In this Project the methodology/ Approach we used are:

1. **Data Profiling Approach:**

In this approach, data is analyzed to identify its quality characteristics such as completeness, accuracy, consistency, and integrity. This information can then be used to develop data quality rules and metrics to evaluate and improve data quality.

By using this Approach, we focus on improving the Data Quality dimensions on Completeness.

After understanding and analyzing the data we further proceed with Data

**2. Governance Approach / rule-based Approach:**

Data governance approach: This approach involves the implementation of policies, procedures, and standards to ensure data quality. This can include establishing data quality rules, defining data ownership and stewardship, and implementing data quality monitoring and reporting processes.

In this Approach we improvise the DQ dimension of validity based on rule-based approach by defining few rules and standard formats by conforms to the syntax it ca be (format, type, range). Further we try to research on adapting the machine learning libraries for automatic prediction of record linkage using few statistical models and classifiers by resulting the confusion matrix and entity resolution models to improve the data quality. The project will also create interactive dashboards to visualize the data and provide valuable insights.

**Project Process:**

The project involves several major steps, it including understanding the business requirements, and based on selecting the attributes and generating the synthetic data and then later used for organization sensitive customer data.

It further flows through Data Profiling and Data Preprocessing by which includes Data cleaning and anomalies and detection of outliers using few metrics of z-score.

Once we select the critical Attributes, we mainly focused on three data Quality dimensions of completeness, validity and Uniqueness.

**Project Flowchart**

Understanding Business requirements

Synthetic Data Generation from mockaroo

Selection of Attributes based on business requirements

Data Preparation and Preprocessing stage

Improving the Completeness for selected numerical, Categorical attributes

Improving the Vality for

SSN, Telephone, Date of Birth

Improving the Uniqueness

by linking records and deduplication

**Data Quality Issues:**

The project focuses on three dimensions of data quality: completeness, validity, Uniqueness. Improving these dimensions will help PiLog Groups to make better business decisions and maintain better customer relationships.

In our Dataset we are focusing on improving the completeness of the data on few critical attributes that is crucial in the customer data for PiLog groups were:

1. Largest Transaction amount attribute,
2. State and city attributes.
3. Email and Name Parser.

Completeness implies to certain attributes should be assigned the values in a dataset.

Three different levels of constraints can be utilized for applying completeness rules to a data set:

* Essential characteristics that demand a value,
* Optional attributes, which may have a value dependent on a set of requirements.
* Inapplicable qualities, such as last\_name name for an individual, which may not have a value.

For incomplete attribute of Latest Transaction amount, we fulfill customer records by retrieving the Past transaction history from the Organizational metadata and then by using the statistical measurement of max and mean values based on the history and replacing the null values with the respected records in the master dataset.

Transaction History Table

Master Dataset

Metadata DB of PiLog Groups

Master table

Chart

Description automatically generatedChart, histogram

Description automatically generated

1. Conclusions: The project aims to improve the data quality of customer data belonging to PiLog Groups. By focusing on four dimensions of data quality, the project will help PiLog Groups to make better business decisions and maintain better customer relationships. The project involves generating synthetic data,
2. State and City Attributes:

For improving the State and city attributes we have used the API that can retrieve state information based on a given zip code by triggering the Zip code to the Endpoint and then replacing in their respective columns of state and city if there is any null values in the Master data.

ENDPOINT = "http://api.zippopotam.us/us/"

The results we got after improving the completeness is shown in the bar charts below.

Chart, histogram

Description automatically generatedA picture containing histogram

Description automatically generated

1. **Email and Name parser:**

For Email and the Name parser we follow rule-based approach. We make sure there is no null in the Complete name column if not then we retrieved the attribute from the PiLog Metadata to the present master data and then parse it to their respective columns (data enrichment). For email, if the customer doesn’t have a email we provide them with a standard format of [firstname.lastname@Pilog.com](mailto:firstname.lastname@Pilog.com).

We can observe the results from the below bar charts with Percentage of null values.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Improvising the DQ dimension- validity:**

Data Quality rules will fall under these categories. Validation means it asserts what needs to be true about the data we use based on our business expectations. To measure the level of data quality based on the data Quality dimensions, the data we monitor regularly subjected by using the defined rules. Conformance can be calculated as the proportion of valid records for rules that are applied at the record level. The number of instances of invalidity can be counted using both rules that apply to a single table or data set (such as those related to uniqueness) and rules that apply to several data sets (such as those related to reference integrity).

Validation of the Data Quality rules is basically done using several scenarios like data profiling, parsing, Standardization to the required format and cleaning the tools. The attributes we focused on the customer dataset were SSN, Mobile number, and Date of birth.

Diagram

Description automatically generatedAn SSN can be used as a unifying key to combine data from many sources. Joining several data sources becomes much more challenging, even impossible if SSNs are invalid. Consequently, identifying invalid SSNs is one of the crucial tasks in processing data using SSNs. This essay outlines a procedure for verifying SSNs. Three sets of digits make up an SSN. The three-digit area number is the first set. The two-digit group number is the second set. The four-digit serial number is in the third set. Often, hyphens are used to denote the separation between these sets of digits, giving the SSN the format 999-99-9999. Base on the Set of rules from SSA we generated a simple code which can validate the SSN of customers records whether it’s a valid or invalid.

Text

Description automatically generated

In processing data with SSNs is to identify invalid SSNs we can download information from the SSA web site is Available with valid ranges for SSN area and group codes and then standardize the SSN based on the US standard format.

Chart, pie chart

Description automatically generated

Above Pie chart is obtained from Mat plot library from Python which represents the percentage of valid, Invalid, and null values and then we standardize the SSN.

Similarly, we standardize the mobile number to US format by applying some validation rules. Then Final we get into the validation rules for Date of Birth, and we set the rules for the customers above 1900 if the customer data is having an invalid date pf birth then it will set the limits based on the limitations of day/month and year and standardizing to the US format of mm/dd/yyyy.

Graphical user interface, text, application, email

Description automatically generated

The last data quality dimension we focused on is Uniqueness.

Uniqueness refers to the requirements that entities the modeled data within the organization are captured and represented uniquely for every record with the key. And ensures that no entity exists more than once belongs to the same entity.

There are two techniques to monitor this dimension.

It entails doing duplication analysis on the data set as part of a static evaluation to see if duplicate records are there, and as part of an ongoing monitoring process, it entails offering an identity matching and resolution service at the time of record generation to find precise or potentially matching records.

Its the most Critical dimension to ensure no duplication of records.

Measured with in the dataset or across the dataset. Data uniqueness also improves Data Governance and speeds up compliance.

Entity Resolution is a strategy for finding records that refer to the same real-world entity in different data sources or within the same data source. In this even the strings are nearly identical. To improve the uniqueness, we are using Record Linkage Package to link and deduplicate the records.

Purpose: To find the duplicates between two datasets. There are two main types of record linkage.

**Deterministic Record Linkage:**

It involves doing **matches and joins** on data to find duplicates.

Easier, Suitable for data having high level of inaccuracy.

**Probabilistic Record Linkage:**

1. Calculates based on the probability that two records matches or not.
2. Further uses ML Algorithms to classify and find matches automatically using confusion matrix.

We proceeded with 3 methods to improve the Uniqueness and compare the results based on Precision and recall and selecting the best statistical model fit for PiLog group. They are:

1. Record Linkage tool kit (Pairwise comparison)
2. Using Oyster
3. Splink

Stages for improvising the Uniqueness with measuring with Record Linkage tool kit are as follows.

1st stage:

1. Preprocess the data which involves in cleaning the data, data Profiling using pandas which gives detailed Statical, and correlations graphs helps in the data preprocessing stage to detect outliers.
2. Adding *phonetic* version of the name using the Soundex method which clusters the similarity based on phonetics.

2nd stage is Indexing and Blocking which includes identifying the candidate pairs (possibility of having matches) for blocking we used Phonetic Initials and even compared with out phonetics. We compared both candidate pairs record linkages on general blocking of exact last name and phonetic last name.

The candidate pairs with phonetic last names gives high candidate pairs of 104 among 58 duplicate datasets while 55 candidate pairs with exact last\_name.

And details are given below.

**Blocking:**

1. **using exact last\_name**

Graphical user interface, text, application

Description automatically generated

**Using Phonetic Soundex Blocking**

**Using Phonetic Soundex Blocking**

Graphical user interface, text, application

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Generated multiindex showing each candidate pair.

3rd Stage is the Comparisons stage in which it measures the difference between each of the record features it determines level of similar in the feature Using the "jar Winkler" algorithm on all the strings.

Results in scaling the similarity between 0 -1 range.

4th stage includes the Classification stage in which it is used for classifying the records whether it is a Match or No-match.

Selecting an arbitrary score and sum up the comparison values.

If (candidate pair is greater than the arbitrary score, then we consider the candidate pair as a match for those records.

By using the record Linkage method, we further calculate the precision and recall.

**Precision:** Simply it is defined based on the matches we found, how may were the right matches?

**Recall:** It is measured based on how may matches were being missed to pair.

Then further I tried Scores that includes the whole number from 0 to 5.

We can see the change in the effect of precision and recall values.

As the score goes higher the precision value is increased it means we are getting less false positives and however the recall decreases.

There should always be a balance between the precision and recall achieving a good accuracy in finding the duplicate records.

These scores include the fraction values and the results we obtained we can probably predict that the optimal best score is somewhere near to higher limit we set i.e., between 4 and 5.

We can even further proceed with finding the matches automatically using the ML classifiers like SVM because we have the numerical values and SVM usually perform really well on classification problems.

From this we can find the confusion matrix and with further tuning for machine learning (splitting the data to two for test and train data we can still achieve higher precision and recall. We can further calculate the false positives and false negatives.

I am attaching the results below for the reference by comparing the results of record linkage model with the Oyster model approach for finding the record linkages.

Conclusion:

**Comparing the results**

Record Linkage method

when score is 0 precision is 0.5192307692307693 and recall is 0.9310344827586207

when score is 1 precision is 0.5192307692307693 and recall is 0.9310344827586207

when score is 2 precision is 0.5346534653465347 and recall is 0.9310344827586207

when score is 3 precision is 0.7397260273972602 and recall is 0.9310344827586207

when score is 4 precision is 0.9642857142857143 and recall is 0.9310344827586207

**Oyster Results**

Total Count:114 Pairs:50 Clusters:64  
  
Total Distinct Records ID values (N) ............... = 114  
---------------------Clusters ----------------------  
Reference Set Clusters (EC) ........................ = 58  
Link Set Clusters (LC) ............................. = 62  
Reference+Link Set Agreement Clusters (AC) ......... = 64  
  
---------------------Pairs -------------------------  
Total Pairs (D) = ((N \* N-1) / 2) .................. = 6,441  
Reference Set Pairs (E) ............................ = 56  
Link Set Linked Pairs (L) .......................... = 54  
Reference+Link Set Agreement Pairs (A) ............. = 50  
True Positive Pairs (TP) ........................... = 50  
False Positive Pairs (FP) = L - TP ................. = 4  
True Negative Pairs (TN) = D - TP - FP - FN ........ = 6,381  
False Negative Pairs (FN) = E - TP ................. = 6  
  
---------------------Rates - (0.0 - 1.0) -----------  
False Positive Rate = FP / (TN + FP) ............... = 0.00062647  
False Negative Rate = FN / E ....................... = 0.10714286  
Accuracy = (TP + TN) / D ........................... = 0.99844745  
Precision = TP / (TP + FP) ......................... = 0.92592593  
Recall = TP / (TP + FN) ............................ = 0.89285714  
F-Measure .......................................... = 0.90909091  
TWi ................................................ = 0.93697902

**2.Oyster Metrics for measuring the Uniqueness:**

Oyster offers a flexible and extensible record linking framework that supports a wide range of algorithms, including deterministic, probabilistic, and machine learning-based approaches as well as different similarity measures, blocking strategies, and feature extraction techniques. Oyster also offers tools for preprocessing data, visualizing data, and analyzing linkage outcomes using ROC curves, precision, recall, and other metrics.

Oyster is a rule-based approach and we set the rules as mentioned below.

<OysterAttributes System="School">

<Attribute Item="first\_name" />

<Attribute Item="last\_name" />

<Attribute Item="gender" />

<Attribute Item="Phone" />

<Attribute Item="email" />

<Attribute Item="Address1" />

<Attribute Item="Address2" />

<Attribute Item="DOB" />

<Attribute Item="mobile" />

<Attribute Item="ssn" />

<!-- -->

<IdentityRules>

<Rule Ident="1">

<Term Item="ssn" Similarity="SCAN(LR,DIGIT,5,KeepCase,SameOrder)" />

</Rule>

<Rule Ident="2">

<Term Item="first\_name" Similarity="SCAN(LR,LETTER,1,ToUpper,SameOrder)" />

<Term Item="last\_name" Similarity="SOUNDEX" />

</Rule>

<Rule Ident="3">

<Term Item="first\_name" Similarity="LED(0.7)" Similarity2="NICKNAME" />

<Term Item="last\_name" Similarity="SCAN(LR,ALPHA,3,ToUpper,SameOrder)" />

<Term Item="mobile" Similarity="SCAN(LR,DIGIT,5,ToUpper,SameOrder)" />

</Rule>

<Rule Ident="4">

<Term Item="first\_name" Similarity="LED(0.9)" Similarity2="NICKNAME" />

<Term Item="last\_name" Similarity="SCAN(LR,ALPHA,3,ToUpper,SameOrder)" />

</Rule>

<Rule Ident="5">

<Term Item="first\_name" Similarity="LED(0.9)" Similarity2="NICKNAME" />

<Term Item="last\_name" Similarity="SCAN(LR,ALPHA,3,ToUpper,SameOrder)" />

<Term Item="DOB" Similarity="SCAN(LR,ALPHA,4,ToUpper,SameOrder)" />

</Rule>

</IdentityRules>

<Indices>

<Index Ident="1">

<Segment Item="ssn" Hash="SCAN(LR,DIGIT,0,KeepCase,SameOrder)" />

</Index>

<Index Ident="2">

<Segment Item="first\_name" Hash="SCAN(LR,LETTER,1,ToUpper,SameOrder)" />

<Segment Item="last\_name" Hash="SOUNDEX" />

<Segment Item="Address2" Hash="SCAN(LR,DIGIT,0,ToUpper,SameOrder)"/>

</Index>

<Index Ident="3">

<Segment Item="first\_name" Hash="SCAN(LR,LETTER,1,ToUpper,SameOrder)" Similarity2="NICKNAME" />

<Segment Item="last\_name" Hash="SCAN(LR,ALPHA,4,ToUpper,SameOrder)" />

<Segment Item="mobile" Hash="SCAN(LR,DIGIT,5,ToUpper,SameOrder)" />

</Index>

<Index Ident="4">

<Segment Item="first\_name" Hash="SOUNDEX" Similarity2="NICKNAME" />

<Segment Item="last\_name" Hash="SCAN(LR,ALPHA,4,ToUpper,SameOrder)" />

</Index>

<Index Ident="5">

<Segment Item="first\_name" Hash="SOUNDEX" Similarity2="NICKNAME" />

<Segment Item="last\_name" Hash="SCAN(LR,ALPHA,4,ToUpper,SameOrder)" />

<Segment Item="DOB" Hash="SCAN(LR,ALPHA,4,ToUpper,SameOrder)" />

</Index>

</Indices>

</OysterAttributes>

1. **Using SPlink :**

We used SPlink to perform record linkage on the two datasets. Splink uses a probabilistic model to assign a probability score to each pair of records, indicating the likelihood that they refer to the same entity. The model considers the similarity between the variables in the two datasets and assigns weights to each variable based on their importance in identifying the same entity. We used the default configuration of Splink, which includes the Jaro-Winkler distance metric for string variables and the Euclidean distance metric for numeric variables. We also used the default parameters for the model, including the learning rate and the number of iterations.

It is a probabilistic record linkage (entity resolution) helps to deduplicate and link W/O identifiers.

Blocking is done using the Pairwise Comparisons.

Linkage Algorithm is based on Fellegi-Suter's statistical model.

It is best fit for data with multiple columns (are not highly correlated)

**Types of data linking:**

1. dedupe\_only, splink used to find duplicates, for single input dataset.

2. link\_and\_dedupe, splink used to finds links within and between input datasets for two or more input datasets.

3. link\_only, splink used to finds links between datasets, but does not attempt to deduplicate the datasets for two or more input datasets.

Diagram

Description automatically generated

**Analyze the distribution of values in your data.**

Columns with higher cardinality (number of distinct values) The skew of values is important.

Chart

Description automatically generated

**Analyze the distribution of values in your data**

Blocking rules based on pairwise matching with using index\_id

Graphical user interface, text

Description automatically generated

Observations:

Number of comparisons :

substr(l.first\_name,1,1) = substr(r.first\_name,1,1) and l.last\_name = r.last\_name: 52

l.last\_name = r.last\_name: 71

l.email = r.email: 216

l.city = r.city and l.first\_name = r.first\_name: 8

**cumulative\_num\_comparisons\_from\_blocking\_rules\_chart**

Chart

Description automatically generated

**Estimate the parameters of the model**

**probability\_two\_random\_records\_match parameter:**

probability that two records taken at *random represent a match*

**The u values :**

proportion of records in each ComparisonLevel for *truly non-matching* records.

**The m values:**

proportion of records in each ComparisonLevel for *truly matching* records

**Estimate the parameters of the model**

probability\_two\_random\_records\_match parameter:

It follows Deterministic rules

Text

Description automatically generated

**Estimate the parameters of the model**

**The m values :**

Estimating the parameter for all of the other columns by blocking (other than first\_name and surname) using EM (Expectation Maximization Algorithm)

**Graphical user interface, text, application, email

Description automatically generated**

**Estimate the parameters of the model.**

**The u values:**

larger the random sample is the best fit for having the more accurate the predictions.

Graphical user interface, text

Description automatically generated

**Estimate the parameters of the model.**

**The m values:**

Estimating the parameter for all of the other columns by blocking (other than first name and surname) using EM (Expectation Maximization Algorithm)

Graphical user interface, text, application, email

Description automatically generated

**Visualizing model parameters**

Chart

Description automatically generated

1. using machine learning libraries and entity resolution models to improve data quality, and creating interactive dashboards to visualize the data and provide valuable insights.
2. Future Work: In the future, the project can be extended to include other dimensions of data quality, such as timeliness and relevance. The project can also be applied to other businesses to improve their data quality.
3. Chart, line chart

   Description automatically generated

**Visualizing model parameters**

Chart, bar chart

Description automatically generated

**Detecting unlikable records:**

Unlikable records are those which do not contain enough information to be linked.

Unlikable records can be found by linking records to themselves.

* if, even when matched to themselves, they don't meet the match threshold score,
* we can be sure they will never link to anything.

Chart, line chart

Description automatically generated

**Matching based on Match weight and match probability**

resulted Matching records = 63

Table

Description automatically generated

**Match weight Waterfall Chart**

Cumulative comparisions for final match score

Chart, waterfall chart

Description automatically generated

**Results:**

We evaluated the performance of Splink on the two datasets using several metrics, including precision, recall, and F1 score. Precision measures the proportion of true matches among the total number of matches, while recall measures the proportion of true matches among the total number of records that should have been matched. F1 score is the harmonic mean of precision and recall and is a good measure of the overall performance of the model. We also used a threshold of 0.5 for the probability score to determine the matches.

The results showed that Splink achieved high precision and recall, with an F1 score of 0.98. This indicates that Splink was able to identify a high proportion of true matches while minimizing false matches. The results also showed that the name variable was the most important variable in identifying the same entity, followed by the address variable and the date of birth variable.

**Conclusion:**

In this project, based on the comparison of 3 methods to improve the uniqueness we perform record linkage on two datasets containing information about individuals. Linkage record based on Soundex blocking gives high precision and recall values of 0.96 and 0.93.

**Acknowledgement:**

I would like to thank my committee chair and faculty advisor, Dr. John R Talburt, for his guidance and support throughout the project. I would also like to thank my sponsoring organization, PiLog Groups, and my sponsoring supervisor and business consultant, Leon Claassens, for their support and cooperation.

References:

1. <https://www.clearpointstrategy.com/blog/data-quality-metrics>
2. <https://blog.insycle.com/crm-data-quality-checklist#12-identify-data-validation-exceptions-that-break-automation>
3. <https://www.ssa.gov/employer/ssnvsHandbook.pdf>
4. <https://pypi.org/project/recordlinkage/>
5. <https://moj-analytical-services.github.io/splink/index.html>
6. <https://www.robinlinacre.com/maths_of_fellegi_sunter/>
7. <https://www.kaggle.com/code/mikeyhogarth/record-linkage-tutorial/notebook#Classification>
8. <https://towardsdatascience.com/performing-deduplication-with-record-linkage-and-supervised-learning-b01a66cc6882>
9. <https://towardsdatascience.com/performance-metrics-confusion-matrix-precision-recall-and-f1-score-a8fe076a2262>
10. <https://bja.ojp.gov/sites/g/files/xyckuh186/files/media/document/informatica_whitepaper_monitoring_dq_using_metrics.pdf>
11. <https://www.ibm.com/analytics/data-governance>