

# Receipt Matching

# Topics

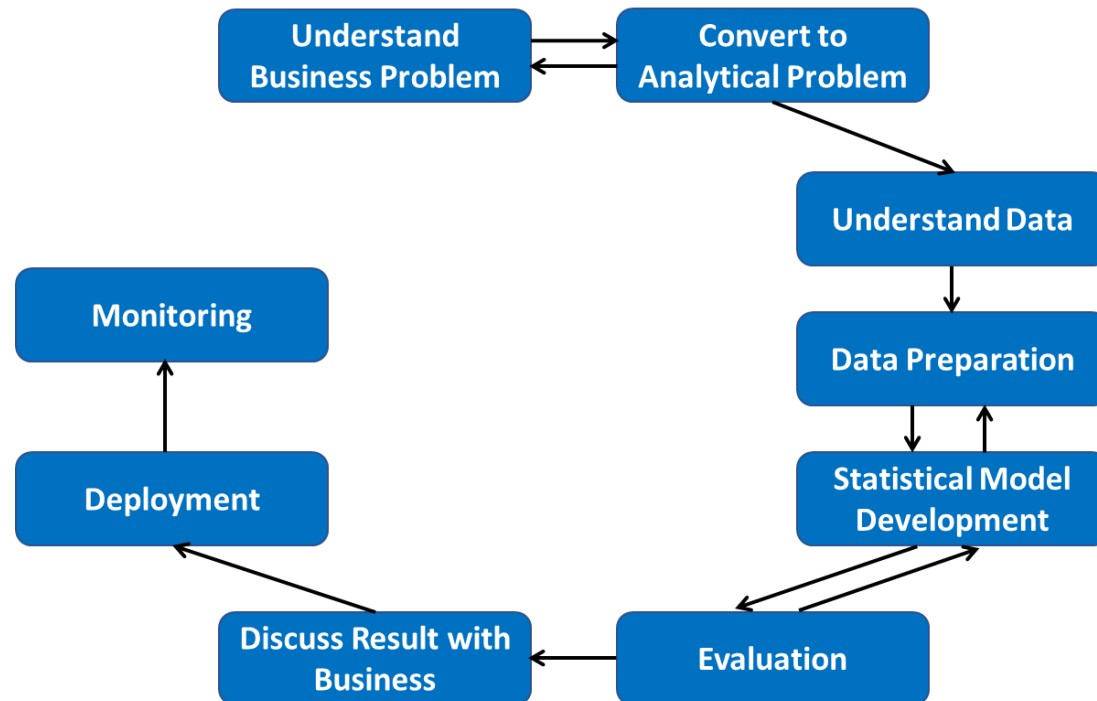
- Business Problem
- Analytical Plan
- Workflow
- Exploratory Data Analysis
- Variable Insights
- Class Imbalance
- Model Development
- Model Performance
- Recommendation

# Business Problem

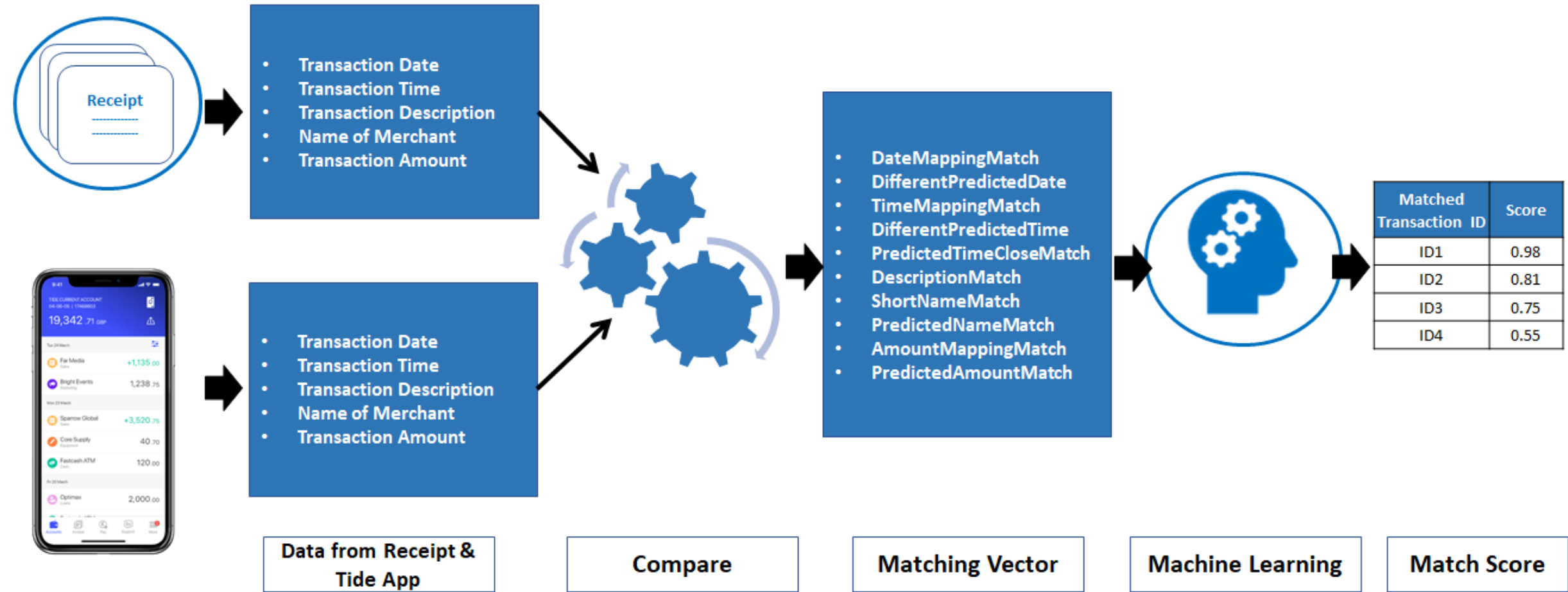


Automatically match receipt images to the associated transaction within the Tide app using data extracted from receipt images and information in Tide transaction.

# Analytical Plan



# Workflow Diagram



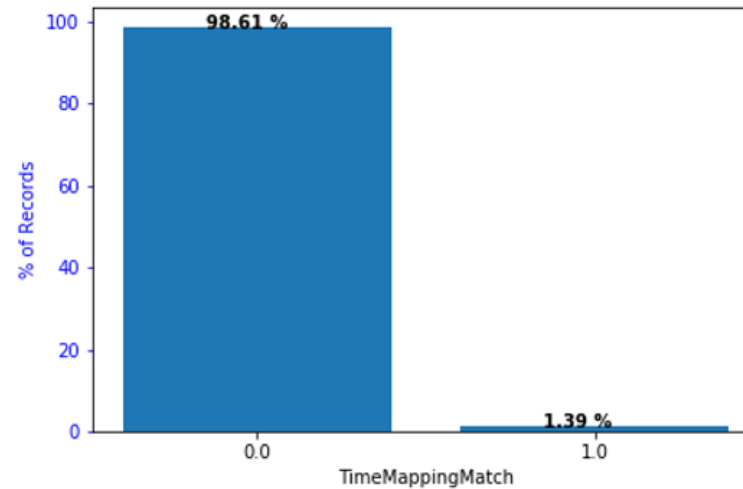
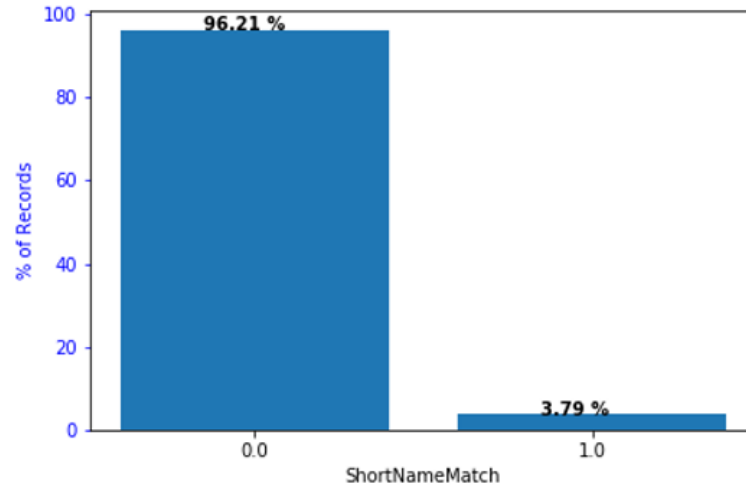
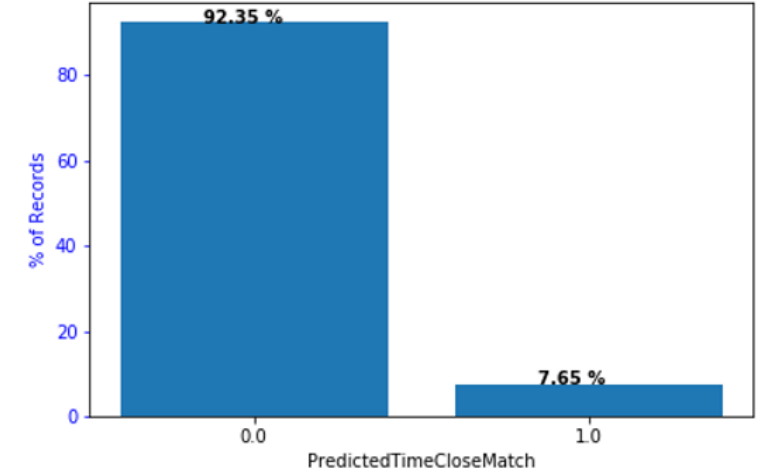
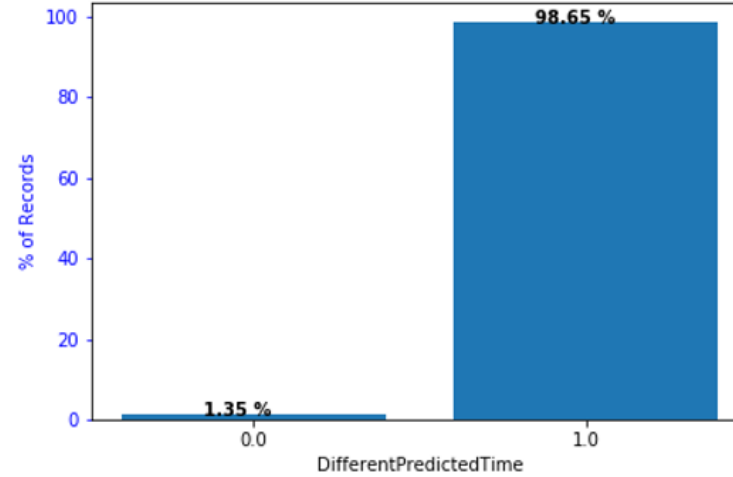
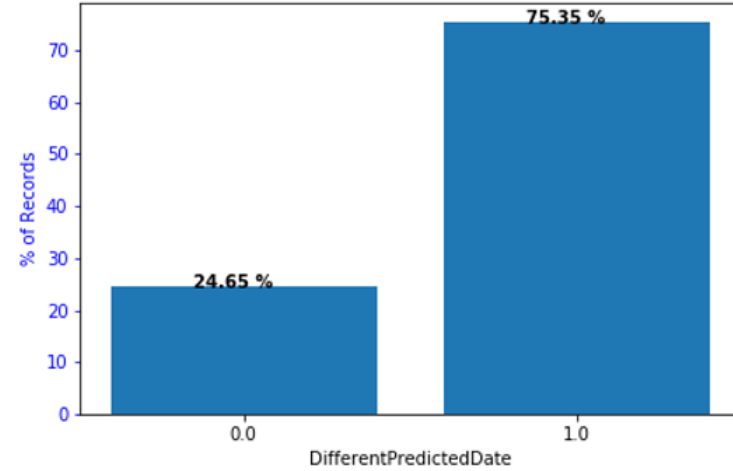
# Data Understanding

The data consists of a csv file with below columns and their descriptions.

Sl No.	Column Name	# Recocrds	# Missing Records	# Unique Values	Data Type	Description
1	receipt_id	12,034	0	1,155	Categorical	Unique identifier for a receipt image
2	company_id	12,034	0	5	Categorical	Tide customer identifier
3	matched_transaction_id	12,034	0	1,155	Categorical	Unique identifier for the transaction that we know is the correct match for the receipt_id
4	feature_transaction_id	12,034	0	2,132	Categorical	Unique identifier for the transaction combined with the receipt_id to produce the matching vector
5	DateMappingMatch	12,034	0	11	Numerical	Matching vector for the given receipt_id and feature_transaction_id
6	AmountMappingMatch	12,034	0	5	Numerical	
7	DescriptionMatch	12,034	0	5	Numerical	
8	DifferentPredictedTime	12,034	0	2	Binary	
9	TimeMappingMatch	12,034	0	2	Binary	
10	PredictedNameMatch	12,034	0	5	Numerical	
11	ShortNameMatch	12,034	0	2	Binary	
12	DifferentPredictedDate	12,034	0	2	Binary	
13	PredictedAmountMatch	12,034	0	6	Numerical	
14	PredictedTimeCloseMatch	12,034	0	2	Binary	

There are 10 Matching vector variables that will be used as input to calculate likelihood score of match between receipt and corresponding transaction in Tide. We will perform Exploratory Data Analysis on these variables to gain more insight about the data.

# Exploratory Data Analysis (Binary Variables)



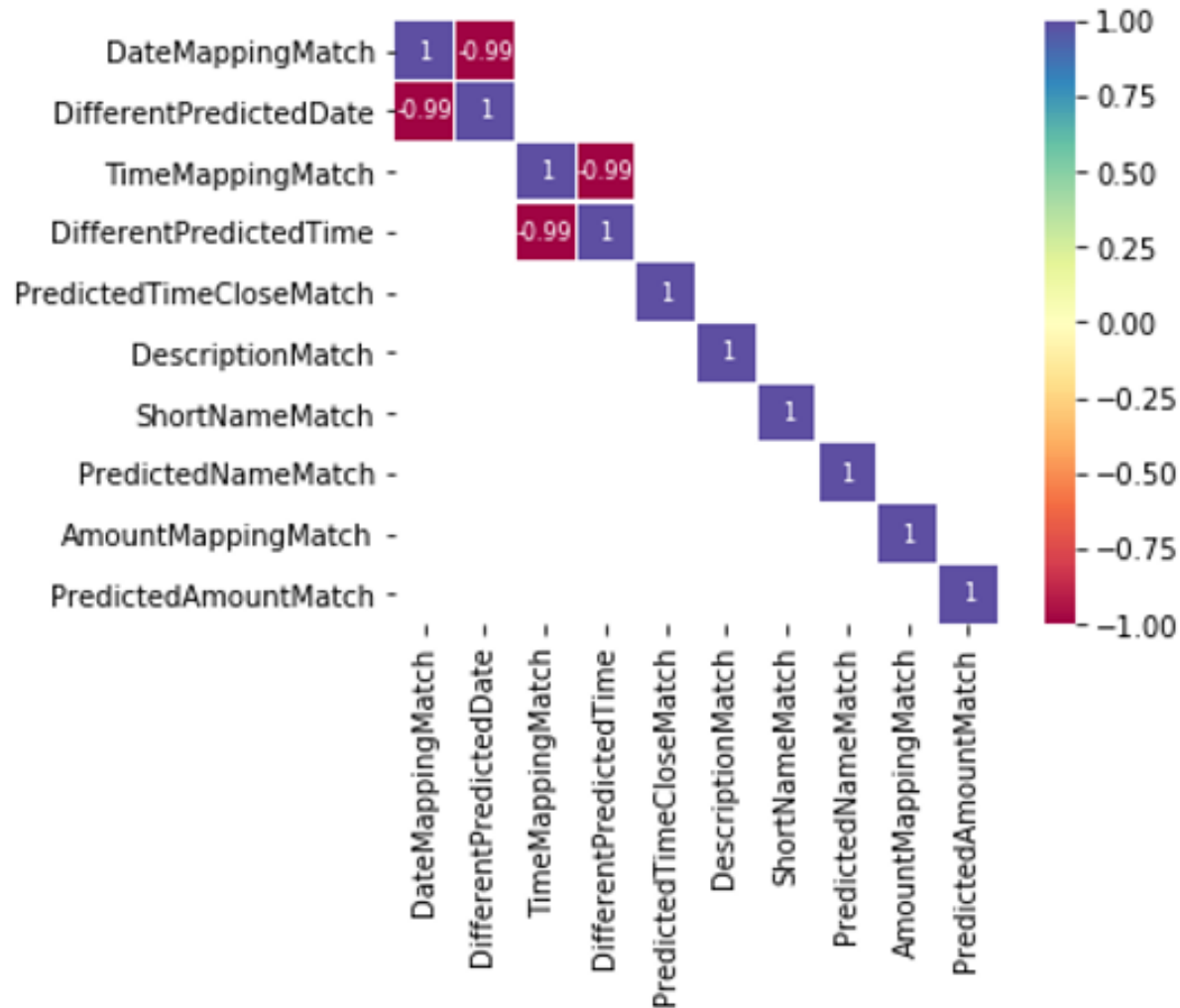
Apart from `DifferentPredictedDate` variable, all other variables are highly imbalanced towards one value. We will examine its behavior in detail with the dependent variable.

# Exploratory Data Analysis (Numerical Variables)

Variable	count	# missing	% missing	# unique value	# zero	% zero	mean	std	min	max	p1	p5	p10	p25	median	p75	p90	p95	p99
DateMappingMatch	12,034	0	0.0%	11	9,068	75.4%	0.2179	0.3845	0	1	0	0	0	0	0	0	0.95	0.95	0.95
AmountMappingMatch	12,034	0	0.0%	5	11,225	93.3%	0.0317	0.1226	0	0.9	0	0	0	0	0	0	0	0.4	0.7
DescriptionMatch	12,034	0	0.0%	5	11,581	96.2%	0.0215	0.1170	0	0.8	0	0	0	0	0	0	0	0	0.8
PredictedNameMatch	12,034	0	0.0%	5	11,589	96.3%	0.0242	0.1286	0	0.8	0	0	0	0	0	0	0	0	0.8
PredictedAmountMatch	12,034	0	0.0%	6	11,989	99.6%	0.0010	0.0201	0	0.6	0	0	0	0	0	0	0	0	0

All the 5 numerical variables are matched scores scaled between 0 to 1 and we have no missing values.

# Multicollinearity



From the correlation matrix it is evident that there is very high inverse correlation between:

- DateMappingMatch & DifferentPredictedDate
- TimeMappingMatch & DifferentPredictedTime

Depending on the type of algorithm that we will use for our solution, we may need to treat this by removing 1 variable from each pair which contains the same information.



# Modeling Base & Target

## Modeling Base

The base on which the model is built is called Modeling base. Here Modeling base is historical data of customers where matching status of receipt with Tide transaction is available along with matching vectors.

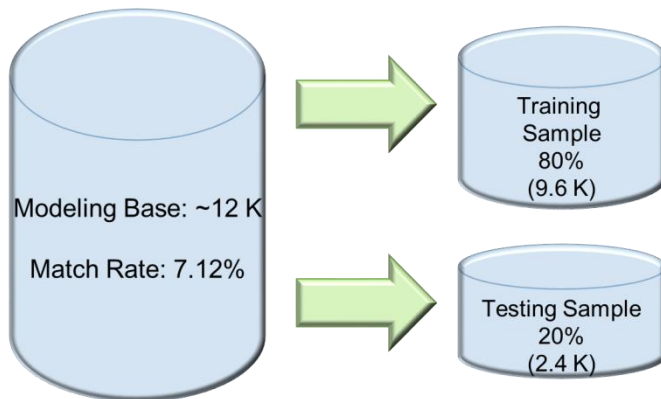
## Target

The phenomenon for which the model is being built is called Target. Here the Target is to predict whether receipt images matches to the associated transaction within the Tide app.

## Target Variable Definition

**Target=1**, if matched\_transaction\_id = feature\_transaction\_id  
**0**, otherwise

Based on this calculation our Target/Match Rate is **7.12 %**



Modeling base is divided into training and testing sample in the ratio of **80:20** to ensure that model is not over fitting on training data and works well with unseen data also.

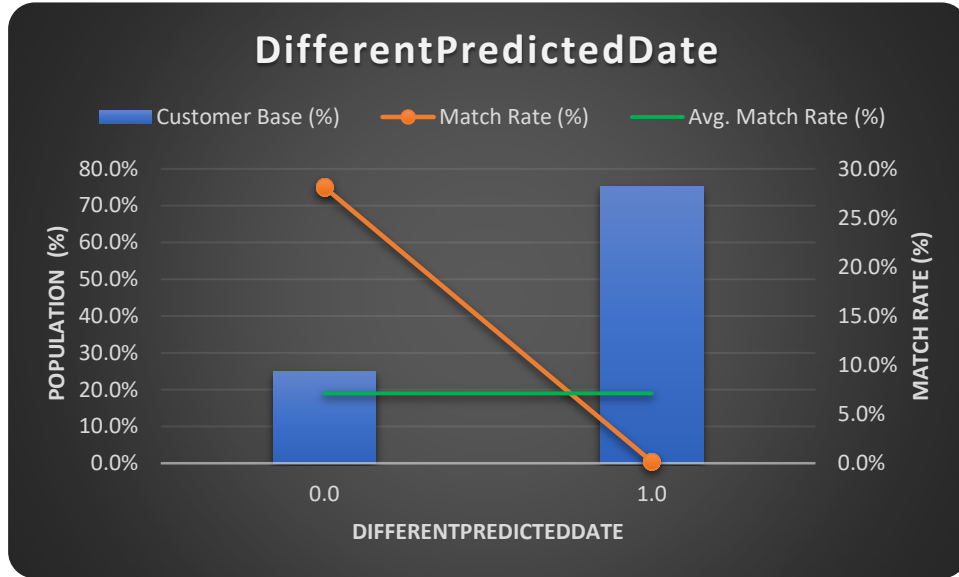
# Information Value

Sl No.	Variable	Information value
1	DifferentPredictedDate	4.71
2	DateMappingMatch	3.41
3	TimeMappingMatch	0.85
4	DifferentPredictedTime	0.84
5	ShortNameMatch	0.65
6	DescriptionMatch	0.63
7	PredictedNameMatch	0.56
8	PredictedTimeCloseMatch	0.44
9	PredictedAmountMatch	0.02
10	AmountMappingMatch	0.02

**Information Value (IV)** helps us to understand the predictive power of independent variables. Interpretation of IV is given in the below table and we can see that 2 of our variables have unexpectedly high IV which is putting them in suspicious category. But after detailed analysis of data it is found that there is no issue with the data.

IV	Information Value Interpretation
<=0.02	Very Weak Predictive Power
0.02 - 0.1	Weak Predictive Power
0.1 - 0.3	Strong Predictive Power
0.3 - 0.5	Very Strong Predictive Power
>0.5	Suspicious/Too good

# Bivariate Analysis with Target

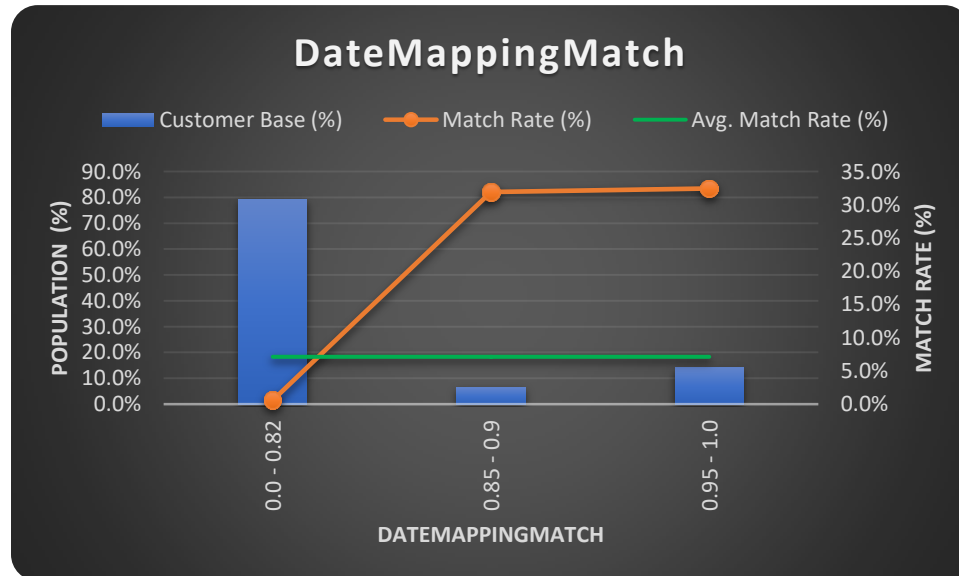


## Nomenclature of this bivariate chart:

**X-axis:** Different segments of the independent variable

**Primary Y-axis:** Population in % for different segments of independent variable

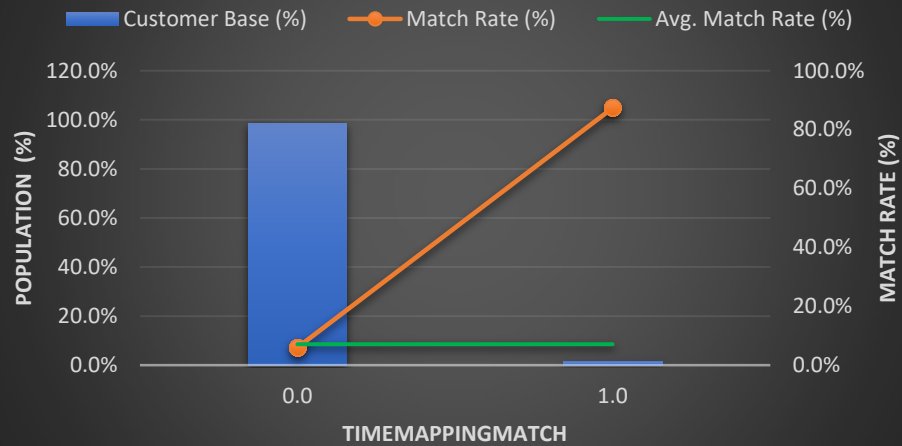
**Secondary Y-axis:** Response Rate in % of different segments of independent variable w.r.t Target variable.



~4X high Match rate in receipt-transaction pair where DateMappingMatch score is greater than 0.82.

As we already know DifferentPredictedDate & DateMappingMatch have high inverse correlation, both the charts are indicating same pattern for match rate.

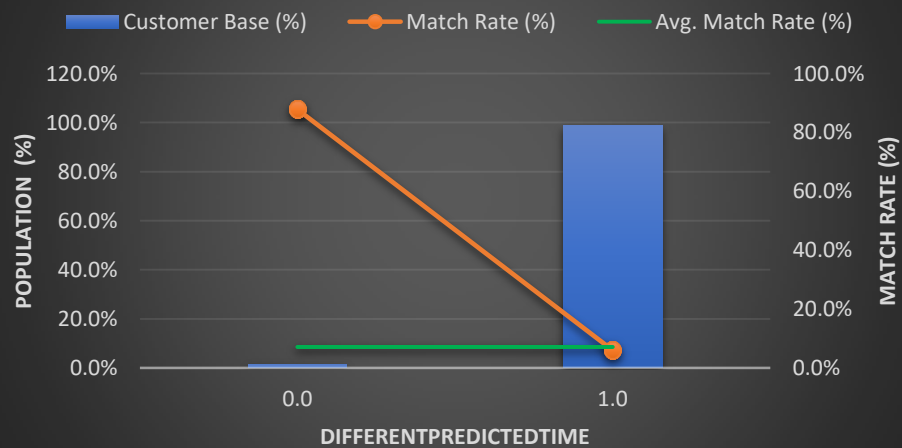
### TimeMappingMatch



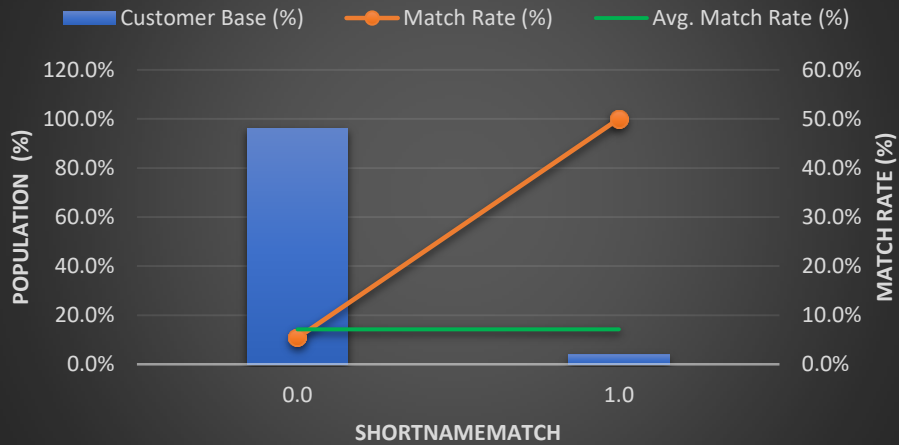
1.4% of receipt-transaction pair have TimeMappingMatch flag=1 and Match rate of this segment is very high ~80%.

This indicates very high predictive power of this variable and as TimeMappingMatch is highly correlated with DifferentPredictedTime variable, it is also showing the same behavior.

### DifferentPredictedTime

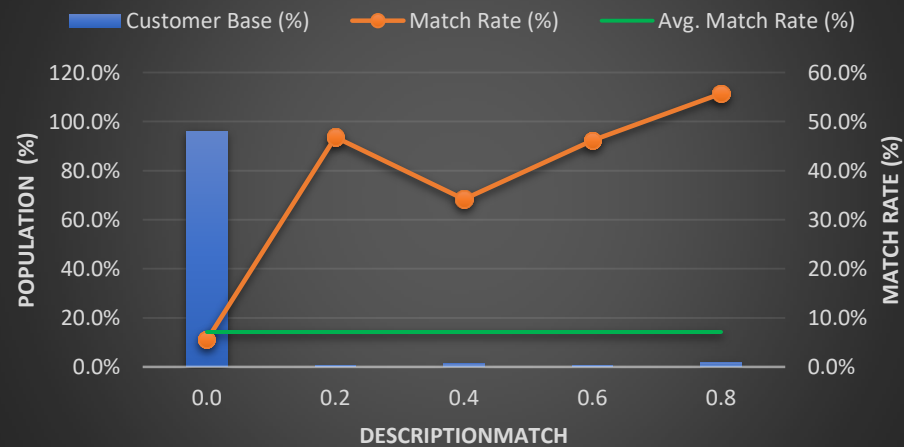


### ShortNameMatch



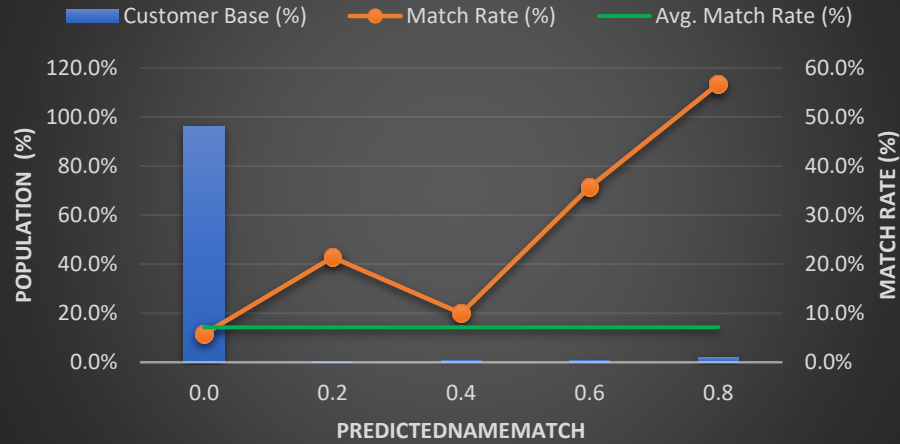
3.8 % of receipt-transaction pair have ShortNameMatch status as 1 and Match rate of this segment is high ~50%. This indicates high predictive power of this variable.

### DescriptionMatch



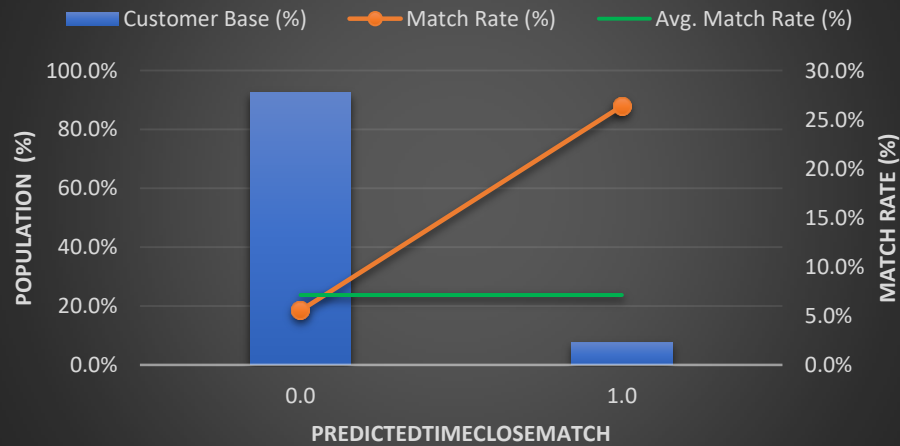
3.9 % of receipt-transaction pair have DescriptionMatch score >-0.2, Match rate is high and varies between 35% - 55%. This indicates high predictive power of this variable.

### PredictedNameMatch



PredictedNameMatch is showing a linear behavior in match rate after score of 0.4 although base size is very small. Match rate is also high for PredictedNameMatch >0.4

### PredictedTimeCloseMatch



PredictedTimeCloseMatch showing a decent lift in match rate ~3.7 X, where this flag is True.

# Class Imbalance

The challenge of working with imbalanced datasets is that most machine learning techniques will have poor performance on the minority class, although typically it is performance on the minority class that is most important.

In our case target variable has a distribution of 7.12 % for match cases and 92.88 % for not match cases.

Oversampling of the minority class is used by us to address class imbalance and our final training data has equal distribution of match and non-match cases.

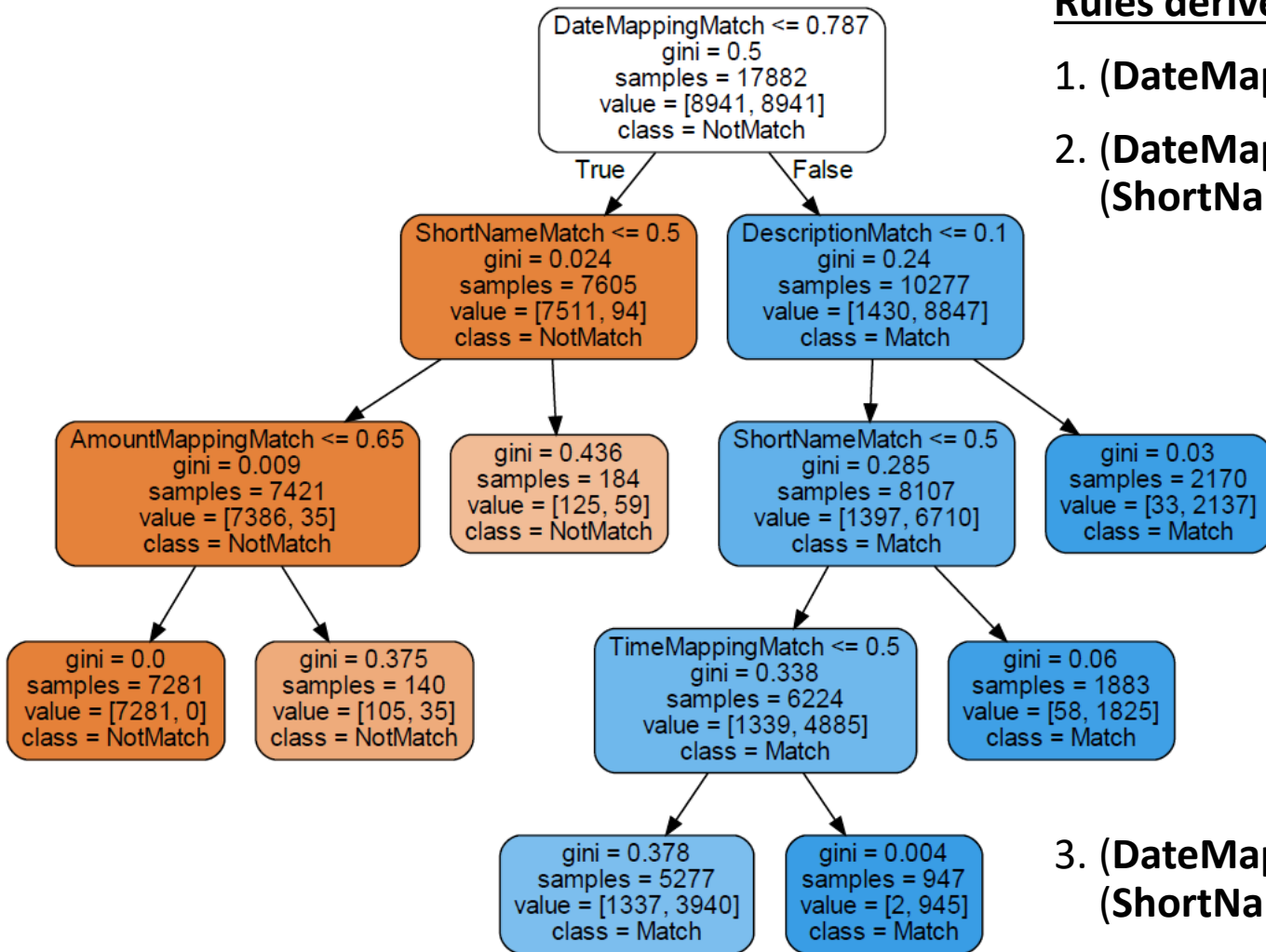
## Model Development

Below 2 techniques are used for model development.

**Decision Tree**

**Logistic Regression**

# Decision Tree



## Rules derived from Decision Tree:

1. (DateMappingMatch ≥ 0.8 ) **AND** (DescriptionMatch > 0.1)
2. (DateMappingMatch ≥ 0.8 ) **AND** (DescriptionMatch ≤ 0.1) **AND** (ShortNamematch == 1)
3. (DateMappingMatch ≥ 0.8) **AND** (DescriptionMatch ≤ 0.1) **AND** (ShortNamematch == 0) **AND** (TimemappingMatch == 1)
4. (DateMappingMatch ≥ 0.8) **AND** (DescriptionMatch ≤ 0.1) **AND** (ShortNamematch == 0) **AND** (TimemappingMatch == 0)



# Logistic Regression

Logistic Regression model summary **statsmodel** python library.

## Logit Regression Results

```
=====
Dep. Variable:          target    No. Observations:          9627
Model:                  Logit     Df Residuals:              9620
Method:                 MLE       Df Model:                  6
Date:                  Wed, 02 Mar 2022    Pseudo R-squ.:          0.5692
Time:                  20:48:14    Log-Likelihood:         -1065.3
converged:              True       LL-Null:                 -2473.0
Covariance Type:        nonrobust    LLR p-value:             0.000
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
const          -7.4170         0.316    -23.467     0.000     -8.037     -6.798
DateMappingMatch  6.2652         0.339     18.505     0.000      5.602      6.929
DescriptionMatch  4.4843         0.345     12.983     0.000      3.807      5.161
TimeMappingMatch  3.7636         0.471      7.999     0.000      2.841      4.686
ShortNameMatch   1.9951         0.172     11.580     0.000      1.657      2.333
PredictedNameMatch 1.4951         0.273      5.477     0.000      0.960      2.030
PredictedTimeCloseMatch 0.5651         0.170      3.318     0.001      0.231      0.899
=====
```

From the model summary of logistic regression, we can see that below 6 variables are statistically significant in predicting transaction-receipt match in the below order of predicting power.

1. DateMappingMatch
2. DescriptionMatch
3. TimeMappingMatch
4. ShortNameMatch
5. PredictedNameMatch
6. PredictedTimeCloseMatch

# Model Performance

Model performance of both Decision Tree & Logistic Regression for training and testing data is given below:

Performance Measure	Decision Tree		Logistic Regression	
	Train	Test	Train	Test
AUC	0.959	0.953	0.962	0.963
True Positive Rate (Sensitivity)	0.987	0.977	0.988	0.988
False Positive Rate	0.160	0.163	0.163	0.167

- For both the models, testing data performance is inline with training data performance. Hence we can infer that both the **Models are not overfitting**.
- Performance of both the models are almost similar (Logistic Regression is slightly better)
- Decision Tree is easy to understand but Logistic Regression gives more granularity to final likelihood score. So for the given problem where multiple possible transaction-receipt matches has to be predicted, it is recommended to use Logistic regression Model.

# Recommendations to improve the outcome

- Possibility of improving the quality of extracted data from receipt should be explored especially for **Description** and **ShortName** variables. Little improvement here can change the final matching output drastically.
- Improving the logic to calculate transaction-receipt matching vector can also help to improve the matching output. Here we can focus on increasing the count of discrete value between 0 and 1 for matching scores.