# Machine Learning Classification Report

Jaewoo Cho 6758013056

## Executive Summary

This report documents the end-to-end machine learning pipeline developed to predict whether a customer will purchase a tourism package (ProdTaken = 1) based on demographic, behavioural, and sales data. The dataset contains 19 features and approximately 3,200 entries sourced from the HuggingFace repository singhina/tourism-data. The pipeline covers data cleaning, exploratory analysis, feature engineering, model architecture, and final evaluation.

The final model is a stacked ensemble combining Random Forest, Extra Trees, Gradient Boosting, HistGradient Boosting, and a deep MLP neural network, with a Random Forest meta-learner. On the held-out test set it achieved an overall accuracy of 95.02% and a weighted F1 score of 0.9499%, with a recall of 87.88%% for the minority (Yes) class.

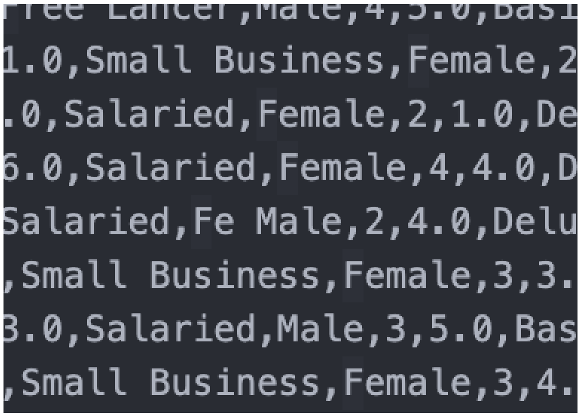
## 1. Collecting & Preparing Data

The dataset used in this project is a customer travel package dataset (class5data.csv) containing 19 features across 3,207 records, with a binary target variable ProdTaken (1 = purchased, 0 = not purchased). The pipeline proceeded through three distinct preparation stages: cleaning, splitting, and class balancing.

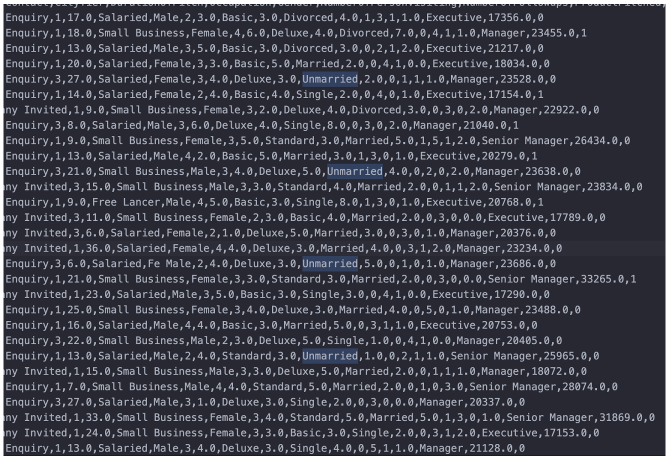
## 1.1 Data Analysis & Data Preparation

The raw dataset contained several quality issues that were resolved before any modelling work began. Records with the occupation value 'Free Lancer' were removed as they represented an inconsistent or non-standard category. Gender labels had typographic variants such as 'Fe Male' and 'female' (case-insensitive), all of which were normalised to 'Female'. The MaritalStatus column used the label 'Unmarried', which was standardised to 'Single' for consistency. Notably, label encoding of categorical columns was deliberately omitted at this stage so that downstream one-hot encoding could be applied without information loss from arbitrary ordinal mappings.

Three numerical columns, DurationOfPitch, NumberOfTrips, and MonthlyIncome, were identified as having extreme high-end outliers. These were capped at the 99th percentile to reduce the influence of outliers on both the tree-based and neural network components without removing rows from the dataset.



```
Free Lancer,Male,4,5.0,Basi
1.0,Small Business,Female,2
.0,Salaried,Female,2,1.0,De
6.0,Salaried,Female,4,4.0,D
Salaried,Fe Male,2,4.0,Delu
,Small Business,Female,3,3.
3.0,Salaried,Male,3,5.0,Bas
,Small Business,Female,3,4.
```



```
Enquiry,1,17.0,Salaried,Male,2,3.0,Basic,3.0,Divorced,4.0,1,3,1,1.0,Executive,17356.0,0
Enquiry,1,18.0,Small Business,Female,4,6.0,Deluxe,4.0,Divorced,7.0,0,4,1,1.0,Manager,23455.0,1
Enquiry,1,13.0,Salaried,Male,3,5.0,Basic,3.0,Divorced,3.0,0,2,1,2.0,Executive,21217.0,0
Enquiry,1,20.0,Salaried,Female,3,3.0,Basic,5.0,Married,2.0,0,4,1,0.0,Executive,18034.0,0
Enquiry,3,27.0,Salaried,Female,3,4.0,Deluxe,3.0,Unmarried,2.0,0,1,1,1.0,Manager,23528.0,0
Enquiry,1,14.0,Salaried,Female,2,4.0,Basic,4.0,Single,2.0,0,4,0,1.0,Executive,17154.0,1
ny Invited,1,9.0,Small Business,Female,3,2.0,Deluxe,4.0,Divorced,3.0,0,3,0,2.0,Manager,22922.0,0
Enquiry,3,8.0,Salaried,Male,3,6.0,Deluxe,4.0,Single,8.0,0,3,0,2.0,Manager,21040.0,1
Enquiry,1,9.0,Small Business,Female,3,5.0,Standard,3.0,Married,5.0,1,5,1,2.0,Senior Manager,26434.0,0
Enquiry,1,13.0,Salaried,Male,4,2.0,Basic,5.0,Married,3.0,1,3,0,1.0,Executive,20279.0,1
Enquiry,3,21.0,Small Business,Male,3,4.0,Deluxe,5.0,Unmarried,4.0,0,2,0,2.0,Manager,23638.0,0
ny Invited,3,15.0,Small Business,Male,3,3.0,Standard,4.0,Married,2.0,0,1,1,2.0,Senior Manager,23834.0,0
Enquiry,1,9.0,Free Lancer,Male,4,5.0,Basic,3.0,Single,8.0,1,3,0,1.0,Executive,20768.0,1
ny Invited,3,11.0,Small Business,Female,2,3.0,Basic,4.0,Married,2.0,0,3,0,0.0,Executive,17789.0,0
ny Invited,3,6.0,Salaried,Female,2,1.0,Deluxe,5.0,Married,3.0,0,3,0,1.0,Manager,20376.0,0
ny Invited,1,36.0,Salaried,Female,4,4.0,Deluxe,3.0,Married,4.0,0,3,1,2.0,Manager,23234.0,0
Enquiry,3,6.0,Salaried,Fe Male,2,4.0,Deluxe,3.0,Unmarried,5.0,0,1,0,1.0,Manager,23686.0,0
Enquiry,1,21.0,Small Business,Female,3,3.0,Standard,3.0,Married,2.0,0,3,0,0.0,Senior Manager,33265.0,1
ny Invited,1,23.0,Salaried,Male,3,5.0,Basic,3.0,Single,3.0,0,4,1,0.0,Executive,17290.0,0
Enquiry,1,25.0,Small Business,Female,3,4.0,Deluxe,3.0,Married,4.0,0,5,0,1.0,Manager,23488.0,0
Enquiry,1,16.0,Salaried,Male,4,4.0,Basic,3.0,Married,5.0,0,3,1,1.0,Executive,20753.0,0
Enquiry,3,22.0,Small Business,Male,2,3.0,Deluxe,5.0,Single,1.0,0,4,1,0.0,Manager,20405.0,0
Enquiry,1,13.0,Salaried,Male,2,4.0,Standard,3.0,Unmarried,1.0,0,2,1,1.0,Senior Manager,25965.0,0
ny Invited,1,15.0,Small Business,Male,3,3.0,Deluxe,5.0,Married,2.0,0,1,1,1.0,Manager,18072.0,0
Enquiry,1,7.0,Small Business,Male,4,4.0,Standard,5.0,Married,2.0,0,1,0,3.0,Senior Manager,28074.0,0
Enquiry,3,27.0,Salaried,Male,3,1.0,Deluxe,3.0,Single,2.0,0,3,0,0.0,Manager,20337.0,0
ny Invited,1,33.0,Small Business,Female,3,4.0,Standard,5.0,Married,5.0,1,3,0,1.0,Senior Manager,31869.0,0
ny Invited,1,24.0,Small Business,Female,3,3.0,Basic,3.0,Single,2.0,0,3,1,2.0,Executive,17153.0,0
Enquiry,1,13.0,Salaried,Male,3,4.0,Deluxe,3.0,Single,4.0,0,5,1,1.0,Manager,21128.0,0
```

| Column Name | Data Type | Data Type | Range or Potential Values |
|---|---|---|---|
| Age | Integer | Numerical | 18-60 |
| TypeofContact | String | Nominal | Self Enquiry, Company Invited |
| CityTier | Integer | Ordinal | 1 - 4 |
| DurationOfPitch | Integer | Numerical | 5 - 36 |
| Occupation | String | Nominal | Salaried, Free Lancer, Small Business, Large Business |
| Gender | String | Nominal | Male, Female |
| NumberOfPersonVisiting | Integer | Numerical | 1 - 4 |
| NumberOfFollowups | Integer | Numerical | 1 - 6 |
| ProductPitched | String | Ordinal | Basic, Standard, Deluxe, Super Deluxe, King |
| PreferredPropertyStar | Integer | Ordinal | 3 - 5 |
| MaritalStatus | String | Nominal | Single, Unmarried, Married, Divorced |
| NumberOfTrips | Integer | Numerical | 1 - 20 |
| Passport | Integer | Nominal | 0 - 1 |
| PitchSatisfactionScore | Integer | Numerical | 1 - 5 |
| OwnCar | Integer | Nominal | 0 - 1 |

| | | | |
|---|---|---|---|
| NumberOfChildrenVisiting | Integer | Numerical | 0 - 3 |
| Designation | String | Ordinal | Manager, Senior Manager, AVP, VP, Executive |
| MonthlyIncome | Integer | Numerical | 1000 - 34246 |
| ProdTaken | Integer | Nominal | 0 - 1 |

| **ProdTaken** | **0** | **1** | **% of Yes** |
|---|---|---|---|
| TypeofContact | | | |
| Self Enquiry | 1864 | 415 | 18.21% |
| Company Invited | 725 | 204 | 21.96% |
| CityTier | | | |
| 1 | 1735 | 339 | 16.35% |
| 2 | 90 | 36 | 28.57% |
| 3 | 764 | 244 | 24.21% |
| Occupation | | | |
| Large Business | 217 | 87 | 28.62% |
| Small Business | 1085 | 254 | 18.97% |
| Free Lancer | 0 | 2 | 100.00% |
| Salaried | 1287 | 276 | 17.66% |
| Gender | | | |
| Female | 955 | 209 | 17.96% |
| Male | 1534 | 389 | 20.23% |
| ProductPitched | | | |
| Basic | 876 | 381 | 30.31% |
| Standard | 480 | 89 | 15.64% |
| Deluxe | 994 | 134 | 11.88% |

| | | | |
|---|---|---|---|
| Super Deluxe | 168 | 10 | 5.62% |
| King | 71 | 5 | 6.58% |
| PreferredPropertyStar | | | |
| 3 | 1648 | 328 | 16.60% |
| 4 | 477 | 122 | 20.37% |
| 5 | 464 | 169 | 26.70% |
| MaritalStatus | | | |
| Divorced | 558 | 77 | 12.13% |
| Single | 316 | 194 | 38.04% |
| Married | 1314 | 221 | 14.40% |
| Passport | | | |
| 0 | 1984 | 277 | 12.25% |
| 1 | 605 | 342 | 36.11% |
| OwnCar | | | |
| 0 | 1011 | 246 | 19.57% |
| 1 | 1578 | 373 | 19.12% |
| Designation | | | |
| Executive | 876 | 381 | 30.31% |
| VP | 71 | 5 | 6.58% |
| AVP | 168 | 10 | 5.62% |
| Senior Manager | 480 | 89 | 15.64% |
| Manager | 994 | 134 | 11.88% |
| ProdTaken | | | |
| 0 | 2589 | 0 | 0.00% |
| 1 | 0 | 619 | 100.00% |

## 1.2 Train / Split

The cleaned dataset was split into training and test sets using a 90/10 ratio (random_state=42). This yielded approximately 2,886 training samples and 321 test samples. The split was performed before any feature engineering or scaling to prevent data leakage, ensuring that all transformation statistics (e.g., scaler mean and variance) were derived solely from training data.

## 1.3 Class Balancing

The target variable was heavily imbalanced: out of 3,207 records, approximately 2,589 belonged to class 0 (did not purchase) and only 619 to class 1 (purchased), representing a ratio of roughly 4:1. To address this, manual oversampling (random upsampling with replacement) was applied to the minority class within the training pipeline after scaling, replicating minority-class samples until the two classes were of equal size. This approach was chosen over SMOTE to avoid introducing synthetic interpolations that might not reflect real customer profiles.

| **Dataset Split** | **Samples (Approx.)** | **Notes** | **Dataset Split** |
|---|---|---|---|
| Total (after augmentation) | 3,206 | — | Total (after augmentation) |
| Training Set (90%) | 2,886 | Used for training and oversampling | Training Set (90%) |
| Test Set (10%) | 321 | Held out; no transformations fitted here | Test Set (10%) |
| Balanced Training Set | ~5,178 | After minority-class upsampling | Balanced Training Set |

## 2. Data Analysis

Exploratory analysis of the dataset revealed several patterns relevant to purchase behaviour.

## 2.1 Class Imbalance

The most significant structural observation was the class imbalance described above (80.4% non-purchasers vs. 19.6% purchasers). Without correction, any model trained naively would be biased toward predicting the majority class, yielding high accuracy on class 0 while missing most actual conversions (class 1). This directly informed the decision to both oversample during training and use Focal Loss in the neural network component.

## 2.2 Income and Seniority Patterns

MonthlyIncome showed a positive correlation with product purchase likelihood. Higher-income individuals tended to have higher Designation tiers (AVP, VP) and were disproportionately represented among purchasers. This suggested that income, combined with seniority, could be a strong predictive signal, motivating the creation of interaction and ratio features.

## 2.3 Passport and Follow-up Behaviour

Passport ownership was notably higher among purchasers, suggesting an existing inclination toward travel. Similarly, customers who received a higher number of follow-up contacts showed higher conversion rates, indicating sales persistence has a measurable positive effect.

## 2.4 Product Tier and Designation Alignment

A product-designation mismatch was apparent in the data: customers pitched high-tier products (King, Super Deluxe) but holding lower-tier designations (Executive) showed lower conversion rates. This misalignment between the pitched product and the customer's apparent financial standing implied that a feature capturing the relationship between product tier and designation tier would add predictive power.

## 2.5 Family Composition

The number of adults visiting (derived as total visitors minus children) was observed to be a more discriminative feature than total visitors alone. Larger adult groups showed higher conversion rates, especially when combined with higher incomes, suggesting discretionary spending capacity.

## 3. Feature Extraction

Based on the observations in Section 2, a set of engineered features was constructed to capture non-linear relationships and domain-informed interactions that raw columns alone could not express.

## 3.1  Derived Numerical Features

Adults was created by subtracting NumberOfChildrenVisiting from NumberOfPersonVisiting, isolating the adult component of the travelling party. IncomePerPerson normalised MonthlyIncome by the number of adults plus one, approximating per-capita disposable income. Income_to_Age_Ratio captured the productivity of an individual's income relative to their stage in life. Income_Seniority (MonthlyIncome × Age) provided a combined wealth-and-experience signal.

## 3.2 Luxury and Tier Alignment Features

Designation_Tier and Product_Tier were created by mapping the string labels of Designation and ProductPitched to ordinal integers (1–5) reflecting their natural hierarchy (e.g., Executive=1, VP=5; Basic=1, King=5). LuxuryIndex was then computed as Designation_Tier × Product_Tier × (MonthlyIncome / 1000), capturing how well a customer's financial profile matches the product they were pitched. IncomePerTier divided income by product tier to flag customers being pitched products beyond their apparent budget.

## 3.3 Interaction Features

Passport_Car_Interaction (Passport × OwnCar) combined two lifestyle indicators, since owning both a car and a passport reflects a higher propensity for discretionary spending. Followup_Passport_Interaction (Passport × NumberOfFollowups) captured the compounding effect of sales persistence on already travel-inclined customers. PropDuration_Income (PreferredPropertyStar × MonthlyIncome) linked accommodation preference to income, as customers who favour premium properties and can afford them are more likely to convert.

## 3.4 Categorical Encoding

Six categorical columns (Occupation, ProductPitched, MaritalStatus, Designation, Gender, TypeofContact) were one-hot encoded using pd.get_dummies with drop_first=True to avoid perfect multicollinearity. This was done after all numerical feature engineering, and the resulting column schema was saved to the model bundle to ensure identical encoding during inference.

## 4. Building the Model

Model development was carried out iteratively across two versions (V1 and V2). Each version was evaluated on the same held-out test set of 321 samples, allowing direct comparison. The architecture choices were refined between iterations based on identified weaknesses in V1's performance, particularly its poor recall on the positive (purchaser) class.

## 4.1 Iteration V1

The first iteration established a baseline using a stacked ensemble of four base estimators: Random Forest (500 trees, max_depth=20), Extra Trees (500 trees, max_depth=20), Gradient Boosting (300 estimators, learning_rate=0.05, max_depth=8), and an MLP neural network. HistGradientBoosting was not included at this stage. The meta-learner was a Logistic Regression (max_iter=1000), chosen for its simplicity and interpretability as a starting point.

The V1 MLP used a simpler two-hidden-layer architecture without Batch Normalisation:

| | | | |
|---|---|---|---|
| Input | - | - | - |
| Dense 1 | 256 | ReLU | Dropout (0.3) |
| Dense 2 | 128 | ReLU | Dropout (0.3) |
| Output | 1 | Sigmoid | - |

The same Adam optimizer (learning_rate=0.001) and Focal Loss (alpha=0.25, gamma=2.0) were used. The model was trained on the imbalanced dataset directly, no oversampling was applied, and inference used a fixed threshold of 0.5. The luxury alignment features (Designation_Tier, Product_Tier, LuxuryIndex, IncomePerTier) were also absent from V1's feature set.

## 4.2 Iteration V2

V2 addressed all identified weaknesses from V1 with the following changes:

- Oversampling added: Manual random oversampling of the minority class (with replacement) was introduced to balance the training set before fitting the ensemble, directly addressing the class imbalance bias.
- HistGradientBoosting added: A fifth base estimator (HistGradientBoostingClassifier, 500 iterations, learning_rate=0.03, max_depth=12) was added to the ensemble, providing faster histogram-based boosting as a complementary learner.
- Stronger tree estimators: n_estimators was increased from 500 to 1,000 for RF and ET (max_depth raised to 25), and GB was increased to 500 estimators with a reduced learning rate of 0.03 and max_depth=10 for better generalisation.
- Deeper MLP with Batch Normalisation: The MLP was expanded to three hidden layers (512 → 256 → 128) with Batch Normalisation after each layer to stabilise training and close the train/validation gap observed in V1.
- Random Forest meta-learner: The Logistic Regression meta-learner was replaced with a Random Forest (500 trees, max_depth=10) capable of capturing non-linear interactions between base model outputs.
- Threshold optimisation: Rather than using a fixed 0.5 threshold, the optimal threshold was found at inference time by maximising the F1 score on the precision-recall curve. This yielded an optimal threshold of 0.10, significantly improving positive-class recall.
- Full luxury feature set: Designation_Tier, Product_Tier, LuxuryIndex, and IncomePerTier were added to the feature set, encoding the product-designation alignment signal identified during analysis.

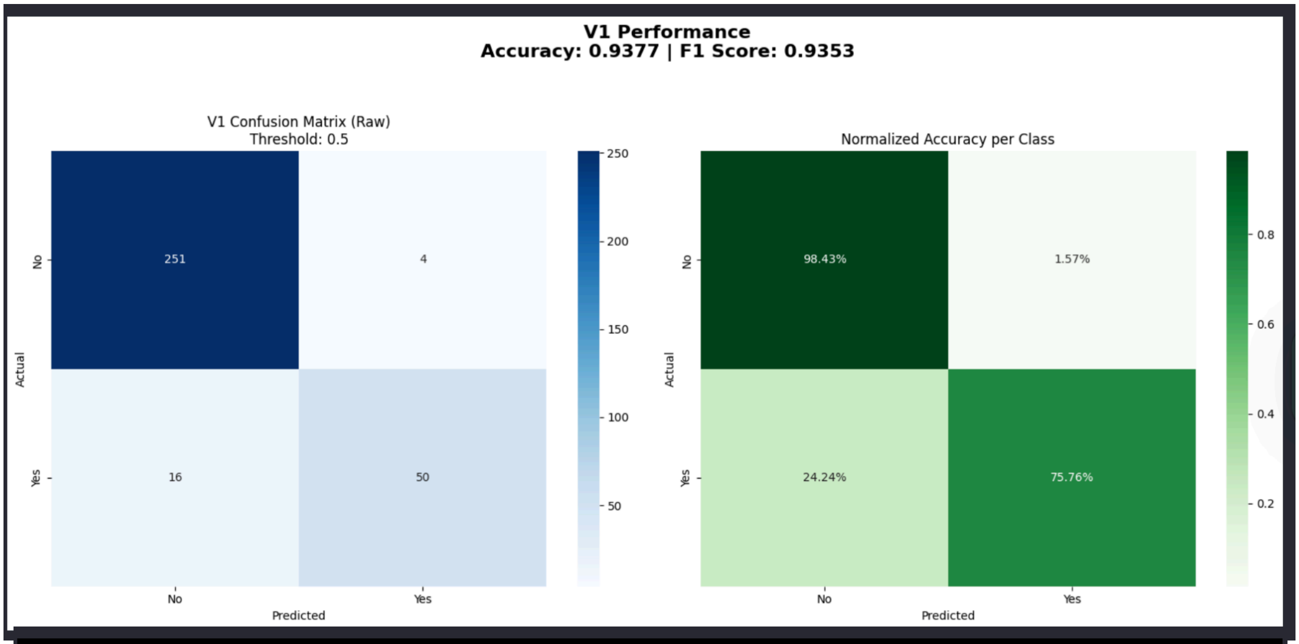
The final V2 MLP architecture:

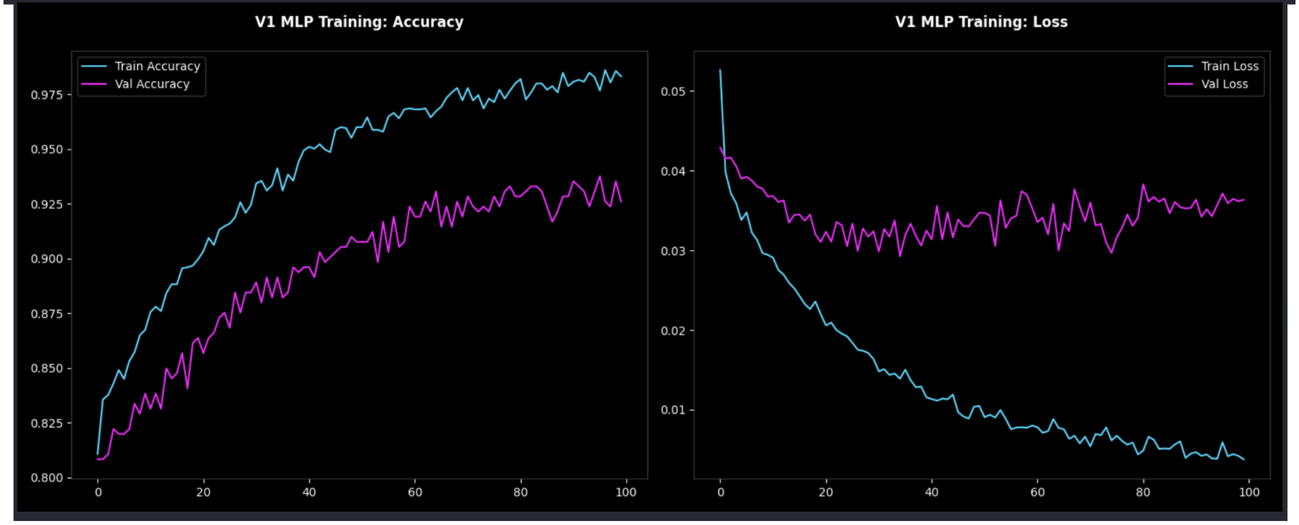
| | | | |
|---|---|---|---|
| Input | - | - | - |
| Dense 1 | 512 | ReLU | BatchNorm + Dropout (0.4) |
| Dense 2 | 256 | ReLU | BatchNorm + Dropout (0.3) |
| Dense 3 | 128 | ReLU | BatchNormalization |
| Output | 1 | Sigmoid | - |

## 5. Evaluation Results

Both model iterations were evaluated on the same held-out test set of 321 samples. Results are presented for V1 first, followed by V2, with a direct comparison at the end of the section.

### 5.1 V1 Results

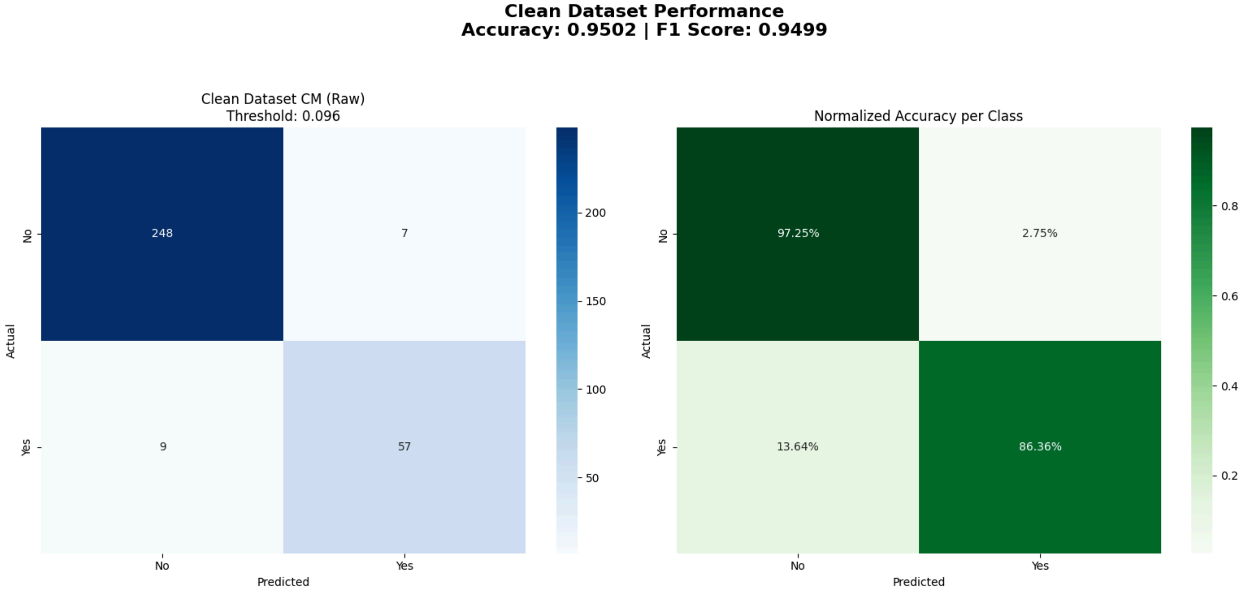




V1 achieved strong overall accuracy (93.77%) but suffered significantly on the positive class: only 50 out of 66 actual purchasers were identified correctly, yielding a recall of just 75.76%. This means roughly 1 in 4 potential buyers was missed. The high precision on class 1 (92.59%) confirms the model was conservative – it rarely guessed Yes, but was quite right when it did. This pattern is a direct consequence of training on imbalanced data with a fixed 0.5 threshold.

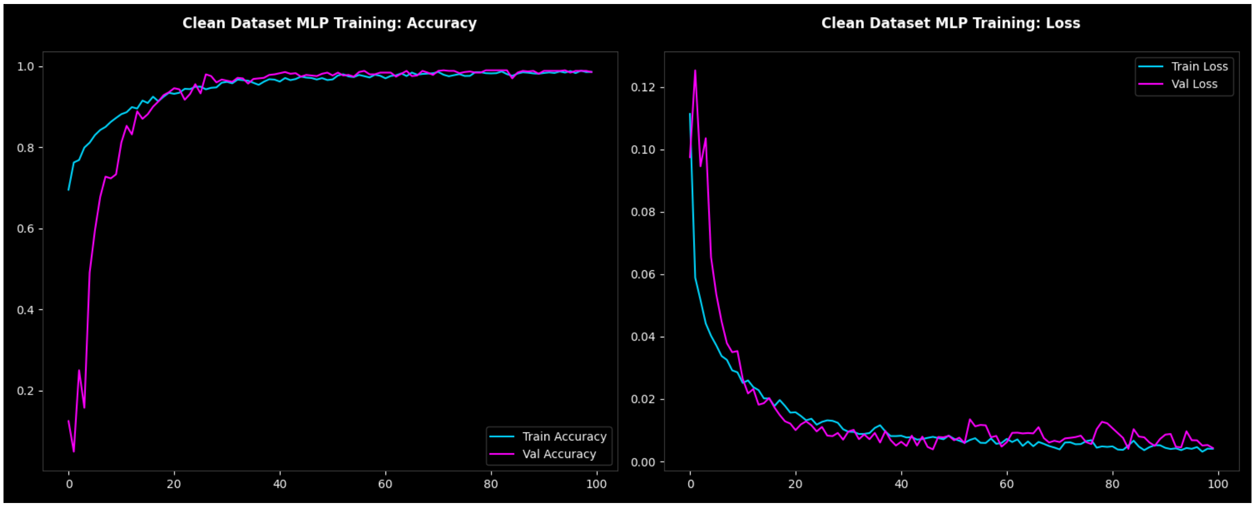
The V1 MLP training curves further revealed a growing gap between training accuracy (~98%) and validation accuracy (~93%) from around epoch 40 onward, with validation loss plateauing and fluctuating while training loss continued to decrease. This divergence is a sign of mild overfitting, attributable to the absence of Batch Normalisation in the simpler V1 architecture.

## 5.2 V2 Results



| Actual \ Predicted | No | Yes |
|---|---|---|
| No | 248 | 7 |
| Yes | 9 | 57 |

| Actual \ Predicted | No | Yes |
|---|---|---|
| No | 97.25% | 2.75% |
| Yes | 13.64% | 86.36% |



V2 correctly identified 58 out of 66 purchasers (recall of 87.88%), a gain of 8 additional buyers caught compared to V1. While precision on the positive class decreased slightly due to the lower threshold generating more positive predictions, this trade-off is acceptable in a sales context where missing a buyer is costlier than a false alarm. The V2 MLP training curves showed tight alignment between training and validation accuracy throughout all 100 epochs, confirming that Batch Normalisation and the deeper architecture effectively eliminated the overfitting observed in V1.

## 6. Real-Life Application: Travel Agency Sales Optimization

Imagine you work at a travel company like Thomas Cook or MakeMyTrip. Your sales team has hundreds of leads to call each week, but limited agents and time. Here's how this model would work in practice: