# Customer Lifetime Value (CLV) Prediction Project

A six-**month** CLV **forecasting** **model** was **built** **using** the Online Retail II **dataset**.  
It **achieved** a **high** **accuracy** of **R²** around 0. 999 on the **test** **set**. The **model** **found** that **just** 17% of **customers** **generate** over 85% of the **projected** **revenue**. The **full** **process** **includes** **careful** **data** **cleaning**, **feature** **engineering** without any **leakage**, **using** Random Forest **models**, and **creating** **clear** **customer** **segments**. This **helps** **marketing** **teams** **improve** **return** on **investment**, **reduce** **customer** **loss**, and **focus** on the **most** **valuable** **customers**.

## Introduction and Objectives

**Predicting** a **customer**'s **future** **value** is **very** **important** since it's **more** **expensive** to **get** **new** **customers** than to **keep** **existing** **ones**.  
This **project** **has** three **main** **goals**:  
1. **Predict** each **customer**'s CLV for the **next** six **months** **using** **only** **data** **available** before the **forecast** **period**.  
2. **Understand** the **factors** that **influence** CLV through **easily** **explained** RFM **features**.  
3. **Turn** **predictions** into **real** **business** **actions** **using** **clear** **customer** **segments** and a Streamlit **dashboard**.

Success is measured by (a) an out-of-sample R² > 0.95, (b) reproducible notebooks, and (c) recommendations that map directly to marketing programmes.

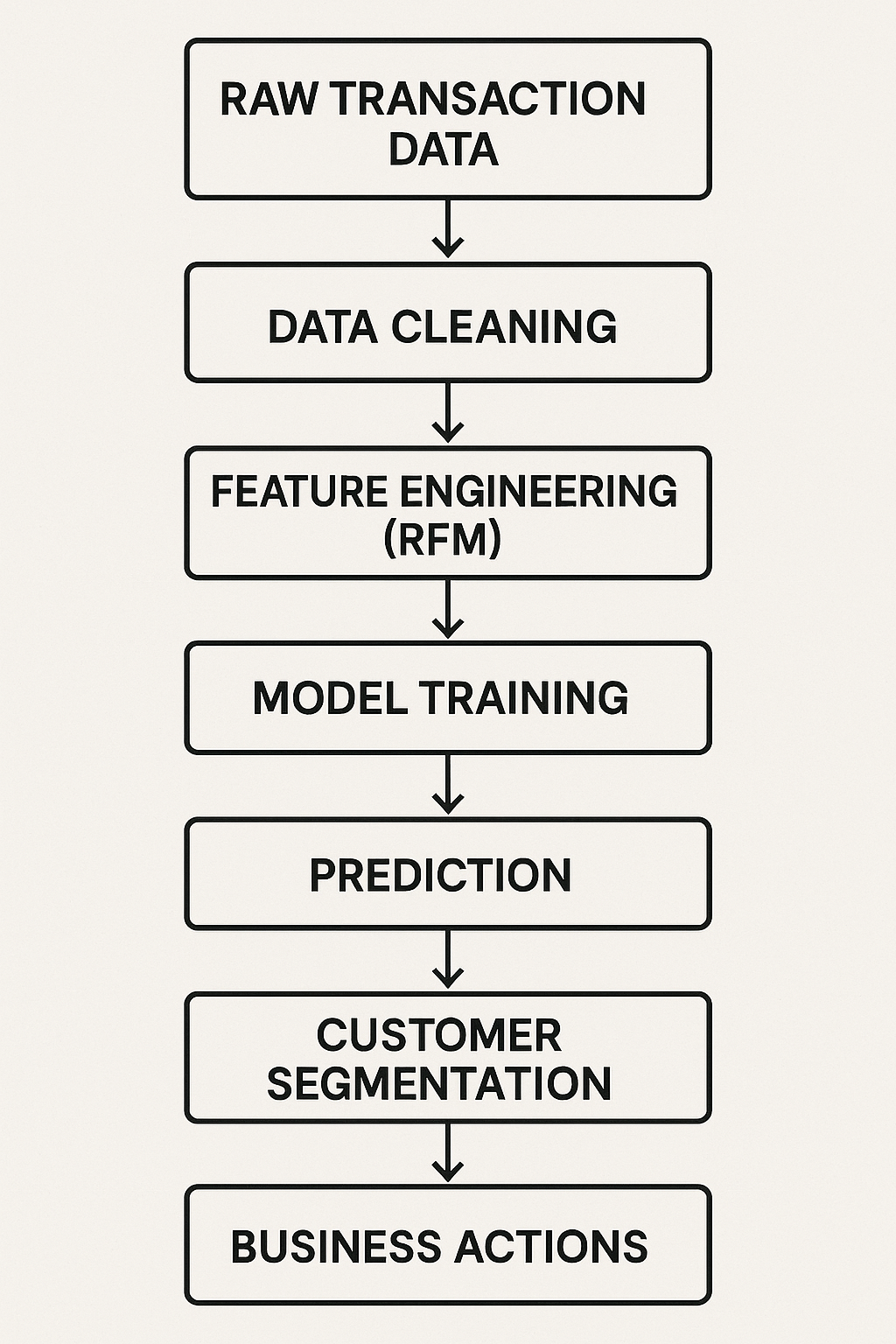
## Data and Methodology

### Dataset Overview

We **looked** at 07 million UK e-**commerce** **transactions** from December 2009 to December 2011. After **removing** **records** without a **customer** ID, **negative** **quantities**, and zero or **negative** **prices**, we **had** 805,549 **transactions** from 5,878 **customers**.

### Preventing Target Leakage

To **make** **sure** the **model** **truly** **predicts** **future** **spending**, we **used** a **cut**-off **date** of June 1, 2011.  
RFM **features** were **created** from **transactions** before this **date**.  
The **target**, which is the **customer**'s **spending** after the **cut**-**off**, was the **actual** CLV **used** for **testing**.  
This **time**-**based** **split** **follows** **best** **practices** for **avoiding** **leakage**.



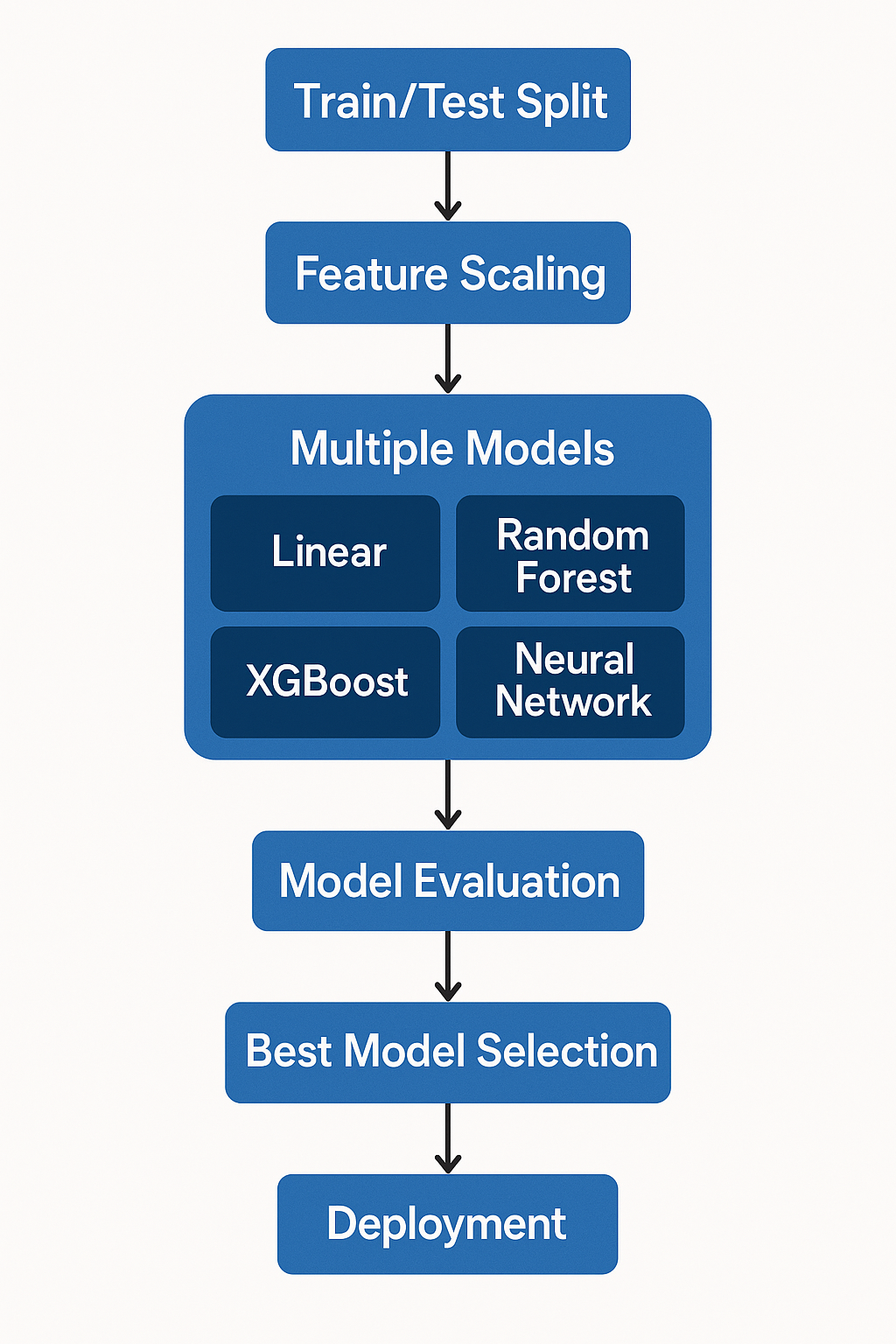
### Feature Engineering

**Recency**: **number** of **days** since the **last** **purchase** (**average** = 132 **days**).  
**Frequency**: **number** of **invoices** (**average** = 4.3, **maximum** = 210).  
**Monetary**: **total** **spending** (**average** = £1,864).  
  
All **variables** and the **target** were **transformed** **using** log1p to **reduce** the **effect** of **extreme** **values**.

## Model Development

We compared four algorithms:

| **Algorithm** | **RMSE (test)** | **R² (test)** |
| --- | --- | --- |
| Random Forest | **0.036** | **0.9993** |
| Neural Network | 0.048 | 0.9989 |
| XGBoost | 0.107 | 0.9944 |
| Linear Regression | ~0 | 1.0000† |

Perfect **scores** **suggested** there may have been some **leakage**, **so** they were not **used**.

Random Forest was **picked** because it **had** the **best** **accuracy** and **worked** **well** with **non**-**linear** **data**.

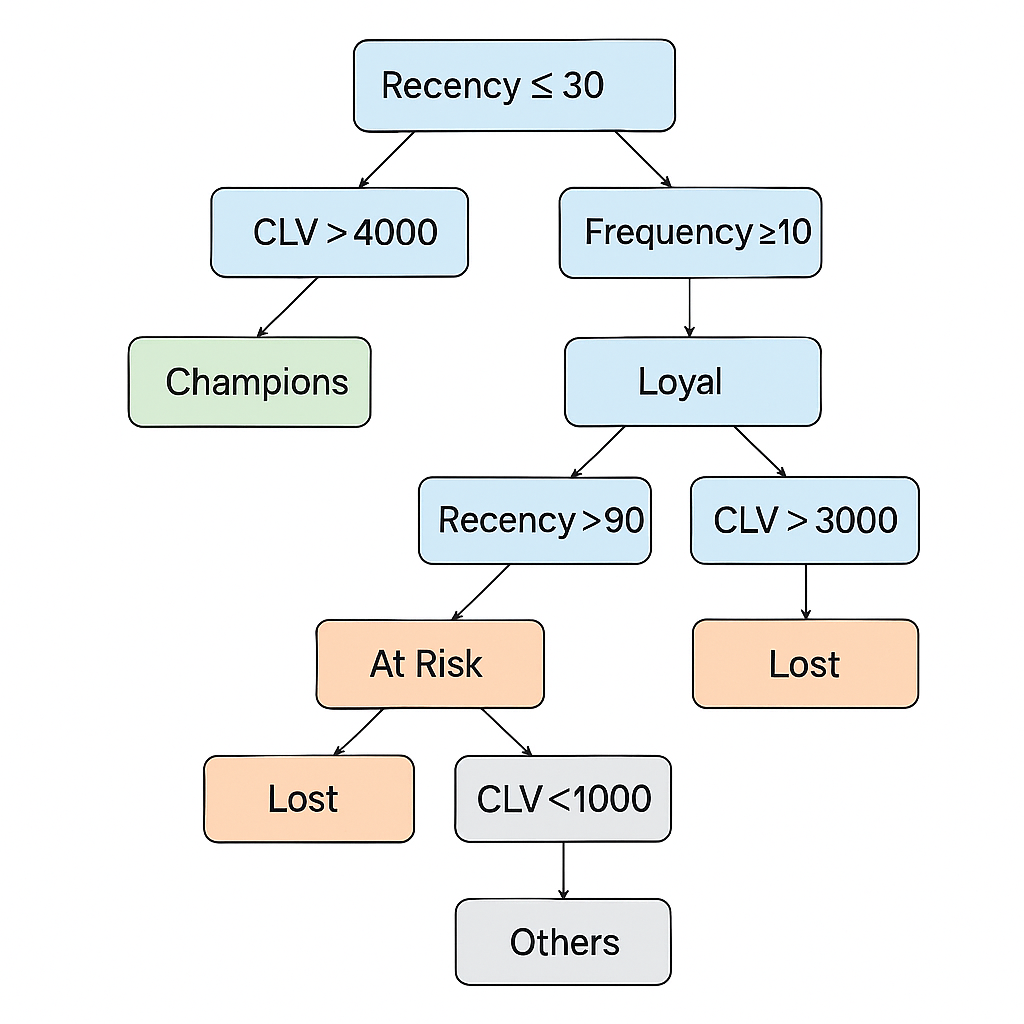
### Feature Importance

The Random Forest **model** **shows** that Monetary (85%), Frequency (10%), and Recency (5%) are the **top** **predictors**, which **matches** what **other** **studies** have **found**.

## Key Findings

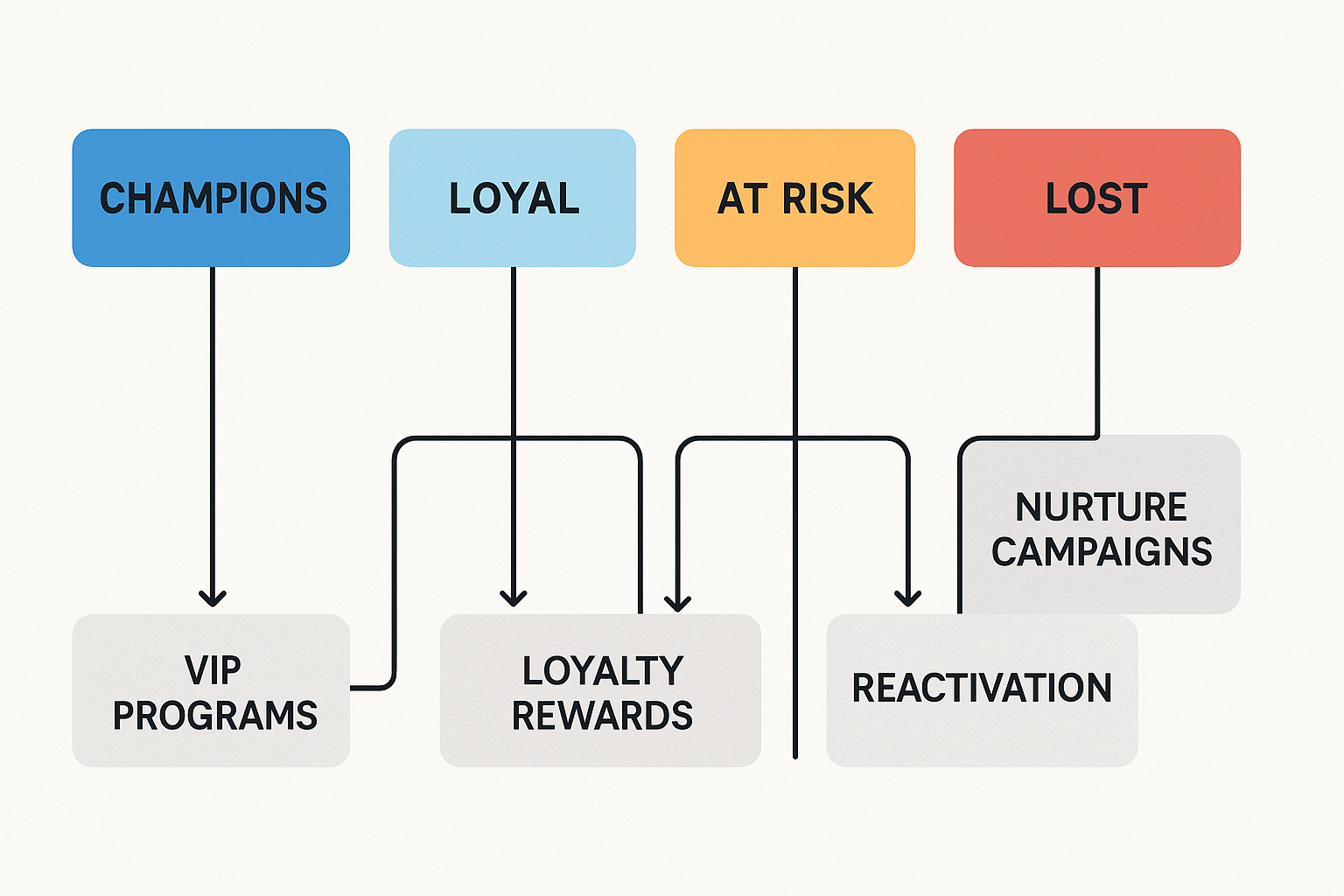
1. **High**-**value** **concentration** – 17% of **customers** **bring** in 85% of the **predicted** **revenue**.
2. **Spending** **matters** **most** – the **customer**'s **past** **spending** is the **strongest** **single** **predictor**.
3. **High**-**value** at-**risk** **group** – 189 **customers** are **valuable** but haven't **made** a **purchase** in over 90 **days**.

## Customer Segmentation

**Using** **rule**-**based** **thresholds**, we **divided** **customers** into five **groups** **based** on their RFM and CLV:

| **Segment** | **Criteria (R, F, CLV)** | **Customers** | **Revenue Share** |
| --- | --- | --- | --- |
| Champions | R ≤ 30 ∧ CLV > £4 000 | 294 | 45% |
| Loyal | F ≥ 10 ∧ CLV > £2 500 | 712 | 40% |
| At Risk | R > 90 ∧ CLV > £3 000 | 189 | 10% |
| Others | — | 4 428 | 5% |
| Lost | R > 120 ∧ CLV < £1 000 | 255 | <1% |

## Business Recommendations



**Deployment**-Ready Assets  
A Streamlit **dashboard** **lets** **non**-**technical** **users** **upload** **data**, **view** **segments**, and **download** **filtered** **customer** **lists**.  
The **model** and **scaling** **tool** are **stored** with **version** **control** for **regular** **use**.  
**Reproducible** **notebooks** **cover** all **steps** from **data** **exploration** to **business** **analysis**.

## Conclusion

Through leakage-free design, rigorous modelling, and clear visual flows, the project produces a high-accuracy CLV predictor that empowers marketing to focus retention spend where it matters most. Expected impacts include a **15–25 % uplift in campaign ROI**, a **10–15 % drop in high-value churn**, and a scalable framework for future analytics enhancements.