**Capstone Project**

**Project proposal**

**1.** **Group description**

**1.1.** Group name

|  |
| --- |
| Team test |

**1.2.** Students names, background and target industry if any

|  |
| --- |
| A: finance  B: graduate  C: retail  → no target industry, open to diverse opportunities. |

**1.3.** Group structure: roles and responsibilities

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Student** | **Data science** | **Project team** | | A | - data engineering  - machine learning  - data visualization | - project proposal submission  - presentation | | B | - data engineering  - exploratory data analysis | - time management  - technical lead | | C | - EDA and insights  - machine learning | - execute project plan  - presentation | |

**2. Why** do we want to develop a data science project?

**2.1 Objective**: what problem do you want to solve? What questions are you trying to answer? How will you **measure the success** of your analysis from a business/user perspective?

|  |
| --- |
| **Context**: the consumer lending department of a retail bank currently takes on average 3 days to approve or reject a car loan, whatever the amount, when the main competitors give an answer in 24 hours for small amounts.  The business approaches the data science team to seek solutions that could be implemented to become more efficient at originating loans while keeping the current accepted level of risk.  **Objective**:  - Predict and assess the level of risk of loan applicants.  - Review and redesign the decision making process in order to automate consumer loan approvals for low risk prospects.  **Measure of success**:  - Decision time for automated approvals < 24 hours.  - Proportion of automated originations for the given level of risk. |

**2.2. Scope** of application: what population and timeframe will your analysis/model be applied to or used for?

|  |
| --- |
| **Population**: process and model will be applied to new customers requesting a consumer loan with an amount below $10,000 (threshold validated with the stakeholders) → ~5,000 requests per month and 1,500 approvals.  For amounts > $10,000, current process will be kept for now.  **Timeframe**: the bank has 10 years of data for this business but the portfolio structure and marketing has critically changed 5 years ago ⇒ we will use the 5 last years of prospects (approved and rejected loans).  **Target variable**: customer default after *m* months of origination.  Questions to answer with business and risk:  - what’s the default definition?  - how many months *m* do we want to consider? |

**3. How** do you translate the objective and scope in terms of data?

**3.1.** What **dataset**(s) do you plan to use? Initial description: source, granularity, number of observations, variables list…

|  |
| --- |
| **Dataset 1: Loan originations data**  - internal system, approved loans  - request level  - 5 years of data → ~90,000  - variables: request id, channel of request, loan amount, term, car details, customer address, FICO score, …  **Dataset 2: Loan servicing data**  - internal system  - transaction level: loan monthly payments  - last 5 years of data → restricted to loans originated in the last 5 years  - variables: loan id, payments, interest rate, term, state, …  **Dataset 3: Bureau data**  - external data  - customer level at time of request  - data stored in monthly .csv files for the past 10 years  - variables: request id, bureau score, number of past requests, missed payments, defaults, etc. |

**3.2.** What **data treatment and analysis** do you plan? Data aggregation, target variable definition, tools, analysis/machine learning, ...

|  |
| --- |
| **Data preparation**  - merge datasets at the loan level → validate ids and keys  - aggregate transaction level data  **Target variable**  - count # of missed payments 3, 6, 12, 18 months after origination  - apply criteria validated with stakeholders  **Tools**  - Data extraction in SQL  - Data preparation in python  - Logistic regression model development in R  - Random forest model development in Python  **Analysis**  - Exploratory data analysis: univariate and bivariate analyses → initial insights to share with stakeholders  - Prediction model: test logistic regression as baseline and random forest |

**4. Project plan**



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **November** | | | | | | **December** | | | | | | | | | | | |
|  | 25 | 26 | 27 | 28 | 29 | 30 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **Kick off** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| *Project declaration* |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data extraction (ETL) |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Data preparation |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Explor. data analysis |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 1** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train logistic regression |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test logistic regression |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Train random forest |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Test random forest |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Finalize results |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Presentation prep |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Buffer |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Milestone 2** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Delivery** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |