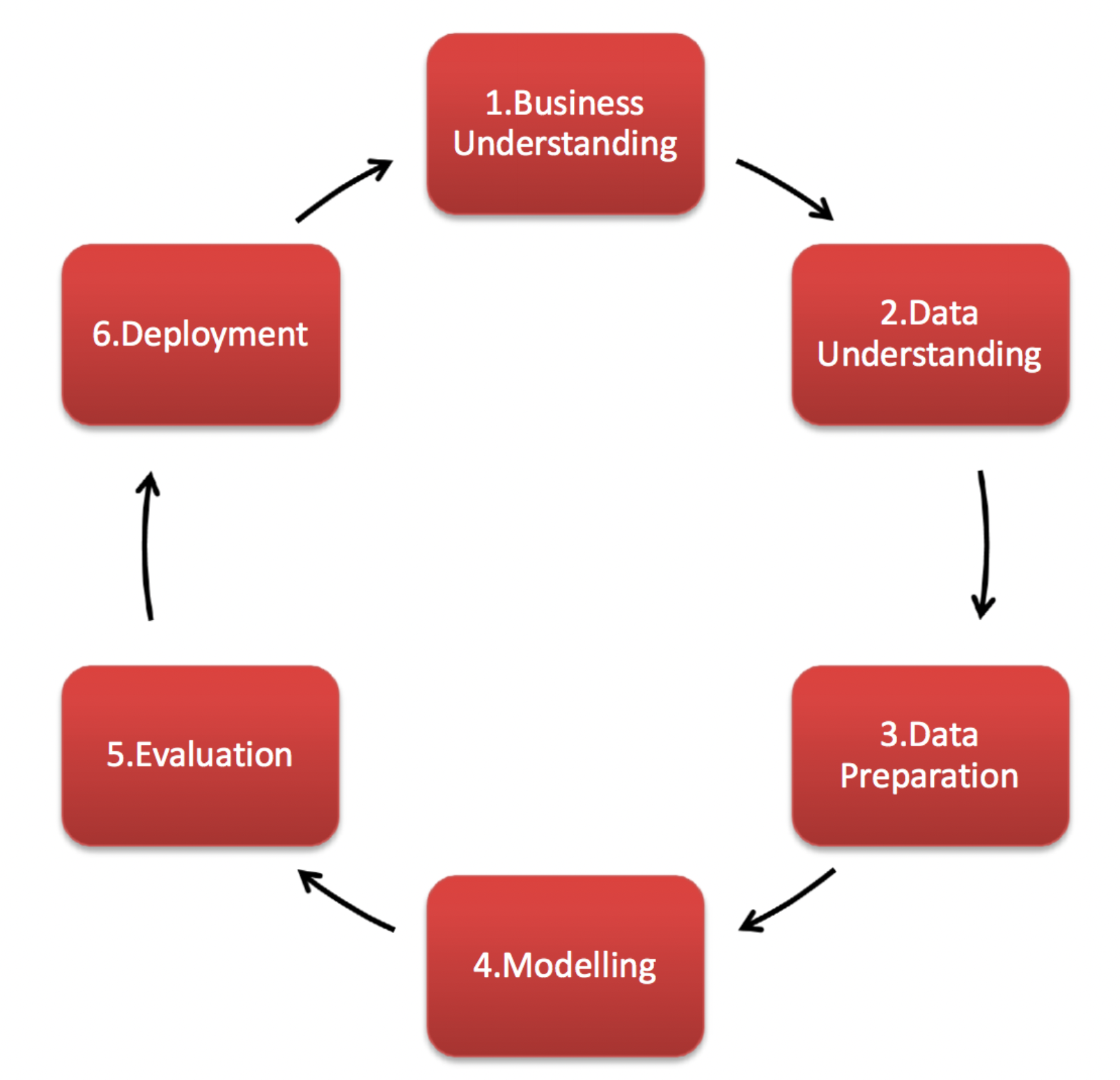
**Part D: Final Report**

The structured approach followed for these experiments from the phase of problem definition to model deployment is CRISP-DM (Cross-Industry Standard Process for Data Mining).



*Fig1: CRISP-DM approach (Patiegm, 2018)*

There are six successive phases in this process (Hotz, 2023):

1. Business understanding – What does the business need?
   1. Business Problem

The main objective is to forecast the likelihood of an existing customer buying a new car with the help of their service record features and other parameters.

The result of this study can be used for the car companies to:

* + - * 1. Develop and execute retention tactics that specifically target customers who are likely to repurchase the car.
        2. Perform better inventory management as the companies would know what models have a higher chance of being resold.
        3. Enhance customer satisfaction by improving certain aspects of car service.

The accurate prediction will give car companies crucial knowledge to focus on the areas that need improvement. The companies can invest their resources to focus on customers likely to make a new car purchase by optimizing their marketing strategies and minimizing unnecessary expenses on ineffective campaigns. They can use this to stay ahead of the competition and improve their business performance.

* 1. Project Goals

The project is carried out through six experiments:

* 1. ***Experiment 1***: Focuses on determining the importance of service in building customer loyalty for car dealerships

1. ***Experiment 2***: Focuses on analysing the top features used to predict the likelihood of an existing customer buying a new car. It involves filtering out the statistically insignificant parameters for the model prediction.
2. ***Experiment 3***: Focuses on predicting the probability of an existing customer purchasing a new car using under-sampled data and logistic regression
3. ***Experiment 4***: Focuses on estimating the likelihood of an existing customer buying a new car. A comparison is drawn between multiple classification techniques to select the most appropriate one.
4. ***Experiment 5***: Focuses on optimizing the hyperparameters of the most accurate model from experiment 4 to get better performance.
5. ***Experiment 6***: Focuses on developing an Explainable Model that can help to understand the contribution of individual features in predicting the likelihood of existing customers buying new cars
6. Data understanding – What data do we have/need? Is it clean?
   1. Data Collection

The dataset for this experiment includes information on customers who have re-purchased a car in the past, along with details on their service history and any other relevant data points. There are seventeen input parameters such as age band, gender, car model, car segment, etc. The dataset includes categorical features only, and out of these features, eleven are divided into ten equal intervals, also known as deciles.

* 1. Data Exploration

The exploration of the dataset involved checking for any missing values, outliers, or anomalies. Also, the relationship between the features and the intended outcome was also studied.

1. Data preparation – How do we organize the data for modelling?
   1. Data Cleaning

Data cleaning is a crucial step in enhancing the model's performance. Handling the missing data was the first step in the data-cleaning process. Each input column that had missing values for more than half of its length was completely eliminated in this case. Whereas inputs with fewer missing values are taken into account when determining the scope for interpolation. The missing values of Gender column is replaced by a third gender as it is a reasonable approach in certain circumstances. In the final stage of data cleaning, textual labels in the dataset were transformed into numerical categories using a label encoder.

* 1. Data Transformation

To ensure that the input parameters are of the same length and magnitude, feature scaling is performed. Feature scaling is done using a technique called standardization which transforms data to have a mean of 0 and a standard deviation of 1. The dataset was split into training and testing data by a 2:1 ratio because this practice helps to evaluate the performance of a machine learning model accurately. To address the imbalanced class issue in the training dataset, an oversampling technique known as SMOTE was applied. This technique involves creating synthetic samples for the minority class to balance the distribution of the classes in the dataset.

1. Modelling – What modelling techniques should we apply?
   1. Model Selection

In the first three experiments, logistic regression was used as a model to predict the discrete outcome variable as it is a simple yet very effective and robust method if the input features are independent and the target variable is having a binary class. In the second experiment, the chi-square test was used to select highly correlated features. For the third experiment, instead of using oversampling, the dataset was modified through under-sampling to evaluate any differences in model performance. This approach involves reducing the size of the majority class to create a more balanced dataset. In the fourth experiment, a comparison is drawn between classification algorithms like Logistic Regression, Naïve Bayes, K Nearest Neighbours, Support Vector Machine, Tree based algorithm (Random Forest) and boosting algorithm (XGBoost). Random Forest gave the maximum accuracy for the training and testing dataset. In the fifth experiment, the hyperparameters of the random forest model were adjusted to improve its performance. The final trial consists of employing boosting algorithms to investigate the significance of features.

Chart, bar chart

Description automatically generated

*Fig2: Comparison of classifier accuracies for testing data*

1. Evaluation – Which model best meets the business objectives?
   1. Model Performance

The model's performance was assessed using accuracy, precision, recall and f1 score. The performance metrics for testing data for each experiment can be seen below in the table.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Experiment** | **Accuracy** | **Precision** | | **Recall** | | **F1 Score** | |
|  |  | **Will not buy** | **Will Buy** | **Will not buy** | **Will Buy** | **Will not buy** | **Will Buy** |
| 1. Logistic Regg. On service features | 77.5% | 0.99 | 0.09 | 0.77 | 0.84 | 0.87 | 0.16 |
| 2. Logistic Regg. On top features | 77.9% | 1.00 | 0.10 | 0.78 | 0.88 | 0.87 | 0.17 |
| 3. Logistic Regg. On undersampled data | 76.3% | 0.99 | 0.09 | 0.76 | 0.85 | 0.87 | 0.16 |
| 4. Comparison of multiple models | 98.6% | 1.00 | 0.81 | 0.99 | 0.85 | 1 | 0.83 |
| **5.Optimization of Random forest model** | **99%** | **1.00** | **0.76** | **0.99** | **0.87** | **0.99** | **0.81** |
| 6. Feature importance and XGBoost | 98.9% | 1.00 | 0.75 | 0.99 | 0.89 | 0.99 | 0.81 |

1. Deployment – How do stakeholders access the results?

Currently, the model is deployed as a Python script, which takes input from the user such as age band, gender, car model, car segment, etc and predicts the likelihood of an existing customer buying a new car. Deploying the final model and making it available for use in production can be done using the following potential steps:

Exporting the model in a pickle file.

Preparation of the data pipeline. CI/CD (Continuous Integration and Continuous Deployment) pipeline is better suited as it focuses on building, testing, and deploying software quickly and reliably. The pipeline also reduces the possibility of bugs and errors entering the production environment.

Deploying the model on a cloud-based platform or a web server. Then, planning the architecture of the containerization and orchestration of the application.

Testing if the model is performing correctly in the production environment.

Monitoring the results of the model over time. This also involves creating a dashboard to keep track of the model's performance and detect any issues.

Improving the model by re-training it over a period of time on the new and updated data.

**Encountered Issues**

Although there is no mention of any sensitive information such as vehicle identification number (VIN) or Personal Identifiable Information (PII) in the dataset, there could still be some data privacy threats such as:

* + - 1. Repurchase data can be utilised for targeted marketing initiatives, which some customers may find intrusive or unwanted. Customers' privacy may have been violated if their information was utilised for marketing without their explicit permission.

**Data Quality Issues**

* + - 1. Small Dataset: The dataset might not be representative of the entire population, so there is a higher chance of bias in the model.
      2. Imbalanced dataset: Imbalanced dataset may be biased towards the majority class (i.e. Customer not buying the car again), resulting in poor performance on the minority class (Customer buying the car again).
      3. High missing values: Some features such as age\_band had really high volume of missing values and had to be removed as they can potentially lead to bias while training the model.

Chart

Description automatically generated

**Modelling Issues**

1. Risk of overfitting: Due to the small size of the training data there was a huge risk of model overfitting.

Features not showing a high correlation: The features selected in the study were not showing a high correlation with the target outcome.

**Future Scope**

There are many ethical issues also involved while applying any machine learning model to a complex task such as:

* + - 1. Bias in the training set
      2. Transparency and Explainability of the model
      3. Privacy and Security concerns
      4. Accountability and responsibility

So, the potential next steps will also involve taking these into consideration.

**Conclusion**

In this project, six experiments were performed and all of them focus on predicting the likelihood of an existing customer buying a new car. The first three involved using the logistic regression model using a set of features and different sampling techniques. The fourth experiment involved comparing multiple classifiers. The fifth trial aimed to optimize the hyperparameters of the most accurate model, while the sixth one used boosting algorithms to determine feature importance. The CRISP-DM methodology was followed to ensure that the project was structured and efficient. It also encourages collaboration between different stakeholders, including business stakeholders, data analysts, and IT professionals. The final model achieved high performance indicating that it can be useful for car companies to develop and execute retention tactics that specifically target customers who are likely to repurchase the car

**References**

Hotz, N. (2023, January 19). *What is CRISP DM?* Data Science Process Alliance. Retrieved March 30, 2023, from https://www.datascience-pm.com/crisp-dm-2/

Patiegm, patiegm. (2018, October 30). *Datasci\_Resources/CRISP-DM analysis template.ipynb at master · patiegm/datasci\_resources*. GitHub. Retrieved March 30, 2023, from https://github.com/patiegm/Datasci\_Resources/blob/master/CRISP-DM%20Analysis%20Template.ipynb