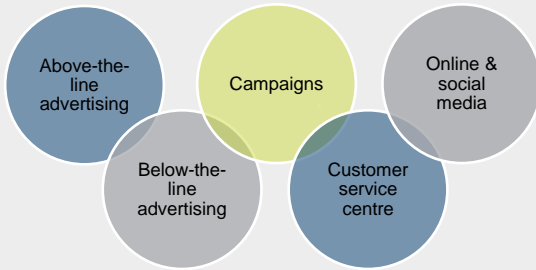# The science of data-driven marketing

Using machine learning to predict which marketing streams Curro should use to reach their target markets better.

1. The Problem

Curro Holdings Ltd is the largest independent schools provider in southern Africa. And currently has 70 407 learners and 76 school campuses, with plans to expand its school network even more.
Curro markets to target customers through five marketing streams:

- Above-the-line advertising
- Campaigns
- Online & social media
- Below-the-line advertising
- Customer service centre

Curro mainly uses geographical market segmentation to decide what marketing content to show to different geographic areas around schools. Marketers decide which marketing stream to market through, by using their intuition and trial-and-error. However, these marketing decisions involve uncertainty. Uncertainty leads to financial losses for Curro in two ways:
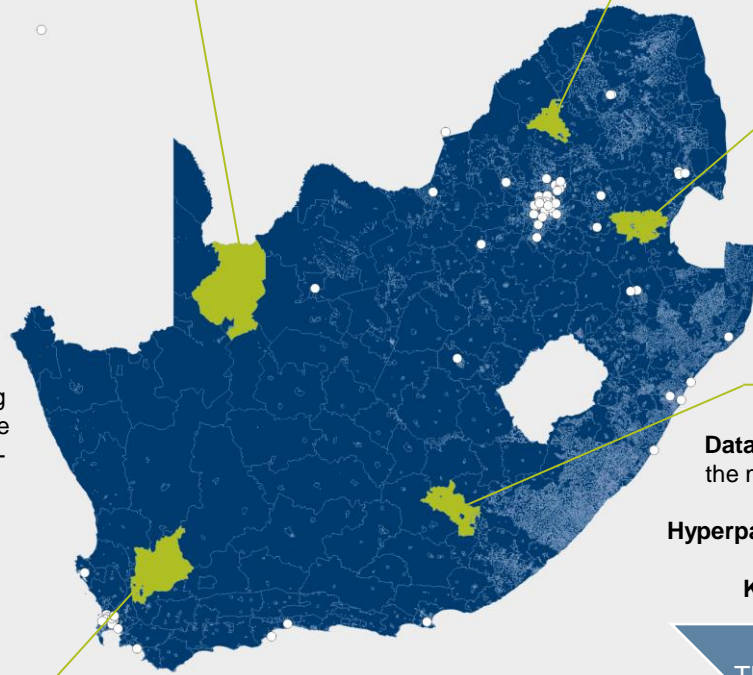- The cost of marketing
- The loss of enrolments

Because school fees are Curro's primary source of income, it is crucial that the right marketing messages reach the right customer through the right marketing stream.

2. The Solution

To develop a machine learning model to predict the marketing stream that should be used in an area based on historical marketing data. The aim is that the machine learning model will classify an input entry into one of the five marketing streams to aid Curro's marketing decision-making and decrease uncertainty in these decisions.

5. Conclusion

The machine learning models successfully predicted which marketing stream should be used in geographic market segments for Curro. The best-performing algorithms were the logistic regression and the support vector machine. Both models can be used to predict marketing streams with nearly 80% certainty. Thus, the machine learning models can support Curro's marketing decisions in new and existing areas.

4. The Results

Machine learning model	Weight F1 score (%)
Logistic regression	79.86
Random forest	76.58
Support vector machine	79.78

The weighted F1 score of the three refined models was compared to determine the best-performing model.

3. Project Approach

Data balancing: The dataset is balanced using sampling methods to remove any bias in the machine learning algorithms towards the majority class in the data, and thus improve the accuracy of the models.

Hyperparameter tuning: Hyperparameter tuning is used to optimise the performance of the machine learning model.

K-fold cross-validation: A technique used to split the input dataset more effectively.

The dataset → Candidate models → Comparison of models → Model refinement

- Age
- Gender
- Race
- Income range

Historical enrolment data was used to train and test the machine learning models.
- Geographical area's data
- Company code
- School brand

Five candidate models were constructed, namely a logistic regression model, a Naïve Bayes model, k-nearest neighbours, a random forest, and a support vector machine.

Candidate models were compared using three classification performance metrics:
- Recall
- Precision
- F1 score