Retention Modelling at Scholastic Travel Company (A)

Machine Learning and Optimisation

MIM 22 Group 23 E2

Alexandra Magdei, Jalel Mohib, Jerry Mao, Margherita Mayr and Rijul Sharma

Executive ummary

Problem

This is a **classification problem** where we need to identify which customers will or will not book trips for next year, which is 2013, based on 2012 historical data.

The business problem is to **craft a tailored marketing strategy** targeting the customer segment and thus enable STC to reduce costs and improve yields.

Models

The two main models that will be considered are **logistic regression** and **classification trees**.

Datasets

Powell will use **STCA_raw_data.csv** as the main datasets after completing the data cleaning process, removing irrelevant variables.

Material handed to Blackford

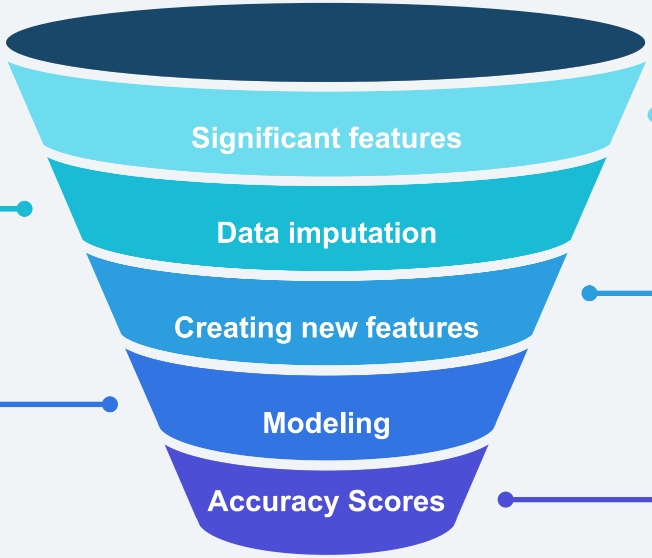
Powell present preliminary results on the **accuracy of each model**, highlighting the best one to predict customers who will book trips.

Data Pre-processing on stca.csv

Problem	Problem identification	Potential Solutions
Missing values	df.isna.any(), this will output all the unique values in the feature. Df.isna.sum()/n this will output the proportion of empty entries	• Use Scikitlearn library to impute • Delete the row/columns where values are missing • Replace the missing value with either the mean, minimum or maximum values. • Run a regression on rows where the data is present and predict for the missing values • Examples: Special.Pay has the majority of missing data; From.Grade & To.Grade has many empty values → drop these
Duplicate rows	Use df.drop(columns=['xxx']) to remove duplicated rows	• Remove duplicate rows
Irrelevant variables & non explanatory variables	After analyzing the relevance of the specific features, one can eliminate the irrelevant ones using df.drop[	• Drop irrelevant variables (ID, Program.Code, From.Grade, To.Grade, Travel.Type, MDR.Low.Grade, MDR.High.Grade, DepartureMonth) • Delete features that are the same for all observations and don't provide additional insight
Incorrect data entry	Use df['xxx'].unique to look at the unique entries that is presented in the specific feature	• Drop entry
Correlated variables	Use a correlation matrix to see if there is any feature with correlation above 0.7	• Remove the feature that has a high correlation.
Mis-formatted data	Dates in string format	• Format correctly: categorical to numerical
Outliers	Compute a box and whisker plot and identify the potential outliers	• Remove outliers or look at what data was intended to be put in the cell

Model Selection Process

This dataset provides **56 features** and **2389 entries**, of which several are irrelevant. Our model selection process is to drop irrelevant features, predict missing variables in specific relevant feature and create new features to gain unique insights for our model.



Which features do not provide any value for our model?

Are there missing entries for relevant features?

Are there features we can combine to gain unique insights?

What type of model would be relevant for this type of dataset?

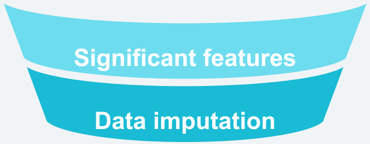
Should Powell use this model:

Yes – high accuracy score

No – redo the model selection process

4

Data Pre-processing on stca.csv

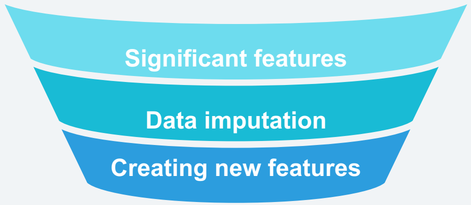


*Our analysis indicates that there are specific features that have missing entries but can be relevant for our model, an example of this is **Poverty.Code**.*

1. Poverty.Code may **indicate the profitability** of each trip. Someone from a lower income level is likely to be more budget constrained and therefore the profitability may be low. Therefore, our marketing strategy and should be taking this into account.

2. Poverty.Code is a "**missing not at random variables"** due to some groups probably not wanting to disclose this private information. For missing not at random variables is much harder to use a data imputation process.

3. We could **run a regression to predict the Poverty.Code**, some ideas may be to divide SPR.Group.Revenue by Total.Pax to get a proxy for revenue per customer and match this with existing data for poverty code.

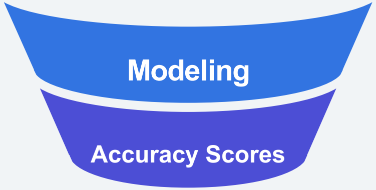


*Our analysis indicates that there creating some new features that combine existing features from the dataset may bear more unique insights, an example of this is **Revenue earned/person**.*

1. Revenue earned per person can indicate which trips are large contributors, this new variable can be used with other features or to impute data as suggested in the last slide's example.

2. This will be done by dividing SPR.Group.Revenue by Total.Pax to get a proxy for revenue per customer. This can help make the data more comparable between each other and therefore bring more insight into trip profitability.

Two Classification Methods



Model Type	Accuracy KPIs	Ways to improve Accuracy	Outcome	Advantages	Disadvantages
Logistic Regression	• Confusion matrix: minimize false negatives • Consider model accuracy • Look at AUC to see how good the model is	• Change the threshold cut off at which students are predicted to book • Use the ROC curve to identify the optimal probability threshold	• Binary outcome	• Good for continuous data types • Good for binary decisions (YES/NO) without further categories	• Assumes that data is linearly or curvy linearly separable in space • Difficult to interpret
Regression Trees	• Confusion matrix: minimize false negatives • Consider model accuracy	• Change the threshold cut off at which students are predicted to book • Use the AUC as a function of nodes to identify the optimal threshold	• The predicted database is divided into multiple leaves	• Easier to visualise and interpret • Non-linear classifiers: do not require data to be linearly separable • Better for datasets with lots of categorical variables • Works well on data with outliers • Better if we need to describe the data	• Involves higher time to train the model • Can cause overfitting • Not suitable for large datasets

Business Insights for STC

1

Business insights from features of programs

STC should analyse the effect that features of programs have on Retention to decide which kind of program STC should invest more resources on and which kind of program STC should focus less on in order to save on costs.
(e.g. Program code, days)

2

Business insights from features of customers

STC should analyse the effect that features of customers have on Retention. This way, STC can create a tailored marketing strategy and relocate its marketing resource to the customers segments that are more likely to retain. It is also efficient to put less resources in segments that has a less possibility to generate income for the firm.
(e.g. school type, Income.Level, School.Sponsor)

3

The Decision Classification Tree

The tree would help identify niche customer segments and create tailored campaigns for them. However, this depends on the granularity of the result that we expect (whether we want to simply predict who will book, or also consider additional features for our marketing strategy).