Understanding Data Assessment

Set: Week 6

Due: 8th August 2022 at noon (11:59am)

Contents

Project description

In Understanding Data, we have stressed throughout the module that effective data analysis leading to insight into the underlying problem is possible, without having to use overly complicated or trendy methods.

This assessment provides you with an opportunity to reveal your own skill in exploring a data-informed problem, building upon the content of the module.

You've been hired by Lu's Communications which is a UK based telecom company that offers the following to its customers:

- Internet
- Landline
- TV

Lu's Communications has a huge problem with churn (loss of customers to competition). It is expensive to acquire new customers and therefore retaining existing customers is much more appealing.

Lu's communications would like to use a targeted approach to identify in advance customers who are likely to churn, which can be then targeted with special programs or incentives. This approach can bring in a huge loss if churn predictions are inaccurate, as money would be wasted to incentivise customers who would have stayed anyway.

Lu's communications have hired **you** to help them with this problem.

You are expected to write a report to management at Lu's communications; these are "non-technical" people.

Related work

Here are some articles regarding customer churn which you **may** find useful.

https://www.superoffice.com/blog/reduce-customer-churn/

https://www.qualtrics.com/uk/experience-management/customer/customer-churn/

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0191-6

Assessment requirements

The structure of the report **must** follow the following structure, or you may be penalised:

- Page 1: Cover page
- Page 2: An abstract of the report
- Page 3: Table of content
- Page 4 to 18: **Main body** of the report where you present your analysis. **You do not need to use all the pages**, but you will be penalised if you go over the page count. Font size of 11 is recommended.
- Page 19: Reference (if applicable)
- Page 20+: An appendix which contains your code. This can be screenshots or code in text format. You will **not be marked on the code**; however, it will be used in case of plagiarism between two submissions. Failure to do this will result in penalisation.

Main body

The main body must contain (in some form):

- Introduction
- Exploratory data analysis
- Initial hypothesis
- Data pre-processing
- Developing and testing machine learning models
- Conclusion

You should also cover the following steps:

1. Data collection
   - Lu's communication has already provided the data (data.csv). However, this data is not perfect: it may contain outliers, missing values and irrelevant (not related to churn) information
2. Data preparation
   - Review the structure of the dataset
   - Peek into the data, summarise your data and get a snapshot of all the features
   - Transform the data into a format ready to apply with the algorithms and modelling
3. Perform EDA (exploratory data analysis)
   - Understand the data and its applicability to the problem
   - Understand how the data and its features are related
   - Evaluate the presence of outliers and their effects
4. Develop an initial hypothesis
   - Based on the EDA, develop an initial hypothesis.
5. Data pre-processing
   - Pre-process that data in an appropriate way. This can include (but not limited to):
     i. Dealing with outliers
     ii. Dealing with missing values

    iii. Dealing with different scales of data
    iv. Feature selection

6. Build and test models
   - Select the appropriate models to predict the churn (regression, classification, etc…)
   - Try at least 4 different models to see which is the best…
7. Conclusion
   - A summary of what you have learnt. This should include (but not limited to):
     i. The best ML model (and its accuracy)
     ii. Comment on your initial hypotheses

Submission

You must submit the report (PDF format) on Aston Blackboard.

Deadline is 8th August 2022 at noon (11:59am).

Dataset

The dataset consist of 7350 observations.

The features of the dataset are:

customer_id: Every customer is given a unique ID

gender: Whether the customer is a male or a female

location: Location of the customer

partner: Whether the customer has a partner or not (1=Yes, 0=No)

dependents: Whether the customer has dependents or not (1=Yes, 0=No)

senior: Whether the customer is a senior citizen or not (1=Yes, 0=No)

Tenure: Number of years the customer has stayed with the company. Lu's communication offers a loyalty scheme; this scheme gives 2% discount on the monthly cost for each year (up to 25 years) the customer remains with the company

monthly_cost: The amount charged to the customer monthly

package: Packages Lu's communication offers. See table below for more information

survey: Score given by customers on the customer service (0=``Poor", 10=``Excellent")

Class: Whether the customer churned or not

Packages

	Package 1	Package 2	Package 3	Package 4
Internet (100MB/S)	✓	✓	✓	✓
Internet speed ad-on (100MB/S)			✓	✓
Internet speed ad-on (100MB/S)				✓
Landline Talks weekends		✓	✓	✓
Landline Talks anytime				✓
TV 100 channels		✓	✓	✓
TV additional 50 channels			✓	✓
TV additional 50 channels				✓
Cost	£26/month	£34/month	£40/month	£ /month

Marking scheme

Report structure and presentation (out of 15)

- [13-15] Clearly presented, all main sections included. Tables, diagrams appropriate and explained. Referencing style correct and consistent.
- [10-12]: Clearly presented, all main sections included. Some tables and diagrams may not be appropriate or incompletely explained. Some referencing but may be inconsistent.
- [5-10]: Parts of report difficult to read or poorly structured. Missing or inappropriate tables, diagrams or references.
- [0-5]: Report incomplete or missing.

Content of report (out of 30)

- [25-30]: Clear and consistent motivation for the approach taken, explanation of the approach, arguments and analysis.
- [20-25]: Mainly clear and mainly consistent motivation for the approach taken, explanation of the approach, arguments and analysis.
- [15-20]: Attempt to describe the motivation, explanation, arguments and analysis, but unclear or inconsistent in parts.
- [5-15]: Incomplete description of the motivation for the approach taken, explanation. No or limited arguments and analysis.
- [0-5]: Report incomplete or missing significant parts.

Use of appropriate tools and methods for analysis (out of 40)

- [30-40]: Comprehensive and clear description of appropriate methods for exploratory data analysis, pre-processing, models including methods not covered in the module.
- [20-30]: Mainly comprehensive and clear description of appropriate methods for exploratory data analysis, pre-processing, models including most topics and concepts developed in the module.
- [10-20]: Description of the methods for exploratory data analysis, pre-processing and models may be unclear or missing in parts.
- [5-10]: Description of application of some appropriate methods for data analysis.
- [0-5]: Very limited analysis. No or very limited attempt made to apply topics and concepts developed in the module.

Hypotheses and results (out of 15)

- [10-15]: Appropriate hypotheses developed from the initial investigations. Results analysed in terms of their support for or against the hypotheses. Models produce appropriate level of accuracy.
- [5-10]: Hypotheses presented but link with initial analysis is not clear. Results presented but not clearly linked to the hypotheses. Models do not produce the expected level of accuracy.
- [0-5]: Missing hypotheses or results. Limited or no attempt to link results with hypotheses.

Penalise

- [10%]: Page count failure
- [10%]: No proof of code in appendix
- [10 %]: Incorrect data set