Lead & Opportunity Scoring

Group 28:
Bing Lesley Yuan 350274, Xin Li 1029029, Xinhui Lu 965246,
Ziyan Zhao 1020405, Raelene Huang 694488

5 June 2020

Contents

1 Introduction

Education and research is a fast-growing industry in Australia, with the expectation to grow 11 per cent in the next five years [16]. As Australia's leading comprehensive research-intensive university, it is imperative for The University of Melbourne to improve its client relationship management with its constituents accordingly to capitalise the growing demand for collaborations and limited resources. There are two important constituent segments at the University: industry partners and major donors. The university offers each of these constituent segments unique offerings such as research collaboration, consultancy, innovation collaboration and philanthropy/giving [2]. Some of these constituents (donor or industry client) choose to donate to the university or collaborate with the university on an ongoing research project. Although the two constituent segments of the University have unique offerings and target different personnel, the basis of their problem is very similar: Given the broad constituents network the university has, there are often more leads and opportunities than can be managed by University staff.

For the purpose of this project we identified two sets of problems: First, identifying the key attributes that would largely influence the lead conversion, to drive business insights. Second, exploring the concept of lead scoring – scoring each lead in terms of its likelihood of success of conversion into a financial agreement and relationships. Each scored lead should aid University staff to prioritise which leads they spend scarce time and resources to optimise the outcomes for the University in the future.

Lead scoring is a well-known machine learning problem, however it can be particularly difficult in the context of this project for the following reasons: Firstly, lead scoring models needs to be able to adapt to the changes from the evolving university – industry collaborations. Secondly, the amount of data required to create an accurate model will be huge. Traditional lead scoring is only able to use data that the university is able to perceive. This means that the client's behaviour during making decisions may be completely lost in the model [19]. There might be several behavioural data that seems very useful for the model in hindsight, but hard to collect in practice. This project will also explore the attributes in this area.

The data comes from the digital transformation program OneCRM, managed by the university two constituent segments, Industry and advancements. Both sets of data include attributes on the client, historical engagements and lead/opportunity attributes. However, we were unable to acquire the actual dataset in time due to privacy issues, a similar public dataset from Kaggle was used for the purpose of this proposal. The public data describes leads from X Education who offers unique products to potential customers, as do Unimelb. The dataset includes historical lead data, product offerings, consumer engagement activities and consumer profile attributes [6].

2 Relevant Literature

2.1 Lead Scoring

Lead scoring, a well-known subtask of customer relationship management (CRM). explained as calculating and assigning a lead score to a company's contacts, from historical characteristics data of the client or contact behavioural data [20]. Characteristics data and behavioural data characteristics data explains the characteristics associated with potential customers, such as customer type, industry, potential customer age (creation date to present) and sales stage. Behavioural data encompasses contact history and other responsibility levels of the contact [18].

There are several CRM system products in the market that promote their abilities to automate a propensity score for leads, for example, AzureML from Microsoft 's Dynamic CRM and Performiture [1, 3]. The advantage of these products is that it calculates the probability of a successful lead based on both the characteristics and behaviour of data, which is what this proposal is aiming to achieve. However, these calculated scores are heavily based on the assumptions that there are no missing data, and the presence of a perfect accurate full data set required by the CRM, which will be difficult to achieve in reality. Moreover, these CRM systems are generalised to the most common sales attributes and do not include specific features associated with the university such as alumni relations.

2.2 Machine Learning Methods

According to Nygård [19], sales lead scoring could be regarded as a classification problem, so supervised learning methods could be applied, for example, logistic regression, random forest and neural network. Nygård experimented on these methods and pointed out that the performance of random forest was better than others. While other approaches state logistic regression would be the optimal choice for lead scoring because it is highly efficient, and the probability of each label could be calculated directly. Therefore, the optimal model is not limited to one.

Logistic regression is not only a useful tool for lead scoring, it can also be used to assess whether a predictor has a statistically significant relationship to the probability of the outcome [12]. Kuhn and Johnson [14] tested on a grant admission probability model with logistic regression and discovered that the model performed poorly when all variables were included. However, removing some insignificant variables improved the model performance significantly. Similarly, the concept of decision tree models allows us to determine the correlation between each model attribute and their average as the basis of predicting the output [13].

Apart from model selection, identifying the most useful features is also a crucial part of this analysis. There are several criteria one should follow when deciding if a feature should be included in the model or not [19]. First, identify

the correlation between features and the label. A feature appears to be more useful if it has higher correlation with the label. Second, if a feature has many missing values, it might be a good idea to exclude it for a large dataset or have the missing value replaced. Lastly, features where most values are different or variables where most values are identical should also be considered.

2.2.1 Missing Data

Handling missing data is one of the most challenging tasks in data analysis because it happens very often and it affects the robustness of data models. Inducing the missing values by eye (deductive imputation) is the most straightforward method. However, even though it can be accurate, deductive imputation cannot be done computationally or applied to large datasets. The most common (and often very good) approach is to replace the missing values with the mean (k-means), which is better for numerical values and proven useful [17]. Chopra [11] proposed a combined approach of imputation by regressing the missing values using random forest, which has been proven useful. It treats missing values as labels and uses the other attributes to estimate the missing labels and evaluated by cross validation.

2.2.2 Data Re-sampling & Evaluation Metrics

The set used to train the model is called the "training" data set and the data set used to estimate the model's performance is called the "test" or "validation" set. To avoid the notion of over-fitting, a common resampling technique is k-fold cross-validation. Performance of each model will then be summarised often by mean and standard deviation to gain understanding of performance [14].

One way to describe the performance of a classification model is through accuracy rate. This is calculated by the correctly predicted over all the instances [7]. Other metrics includes Mean Squared Error (MSE, mean of squared difference between the estimates and the data), precision and recall (which looks at the proportions of true positives identified), and F-Measure (is the weighted harmonic mean of a test). A more common metric is the receiver operating characteristic (ROC) curve, as it can summarize the magnitude of errors of the model performance. This method plots the true positive rate (TPR) against the false positive rate (FPR) across different probability thresholds. The area under the curve (AUC) of the ROC curve is also a useful metric. A higher area under the curve is also indicative of a better model, an incompetent model has an area under the curve of 0.5 [9].

3 Data

3.1 Privacy Impact Assessment

Data privacy and information security have become one of the higher priorities and concerns of many consumers in the recent years due to the rapid development and use of big data and digitalization [4]. Our team was very fortunate to be involved in the process of the privacy impact assessment conducted by OneCRM program team and we were given the opportunity to get to know the process, examine our data, estimate the risk and understand clients impact so we would have deeper understanding on the importance of data security and consequences of data privacy breaches.

3.2 Data Overview

As an initial assessment, we evaluated the features which determine the quality of sales in relation to the dataset of X Education. X Education sells online courses to industry professionals globally, most clients become potential clients by visiting their website. The raw dataset includes 9074 historical leads and 23 features in total, 9 of the features are binary while others are mostly categorial. The data is reasonably balanced with a successful conversion rate of 37.8% among the entire dataset.

Feature	Type	Values	Transforamation
Lead Origin	Categorical	API, Landing Page Submission, etc.	-
Lead Source	Categorical	Google, Organic Search, Olark Chat, etc.	-
Do Not Email	Categorical	0: Do email, 1:Do not email	Factorise
Do Not Call	Categorical	0: No, 1:Yes	Factorise
Converted	Categorical	0: No, 1:Yes	Factorise
Last Activity	Categorical	Email Opened, Olark Chat Conversation, etc.	-
Specialization	Categorical	Business Administration, Media and Advertising, Supply Chain Management, etc	-
What is your current occupation	Categorical	student, housewife, umemployed, employed	-
Tags	Categorical	Will revert afer reading the email, Already a student, Not doing further educaDon, etc	-
Lead Quality	Categorical	High in relevance, Low in relevance, Might be, Not sure, Worset	-
City	Categorical	Mumbai, Thane & Outskirts, Tier II Cities, etc	-
Last Notable Activity	Categorical	Email Spamed, Email Bounced, Email Opened, etc	-
TotalVisits	Numerical	-	-
Total Time Spent on Website	Numerical	-	-
Page Views Per Visit	Numerical	-	-
Fetures Removed and Reason			
Country	Categorical	India, United States, Russia, etc.	Among all of 38 countries, India clients constitute 96.8%.
What matters most to you in choosing this	Categorical	Better Career Prospects, Flexibility & Convenience, Better Career Prospects.	” Better Career Prospects” constitute 99.98%.
Search	Categorical	0: No, 1:Yes	Those sources have almost the same pattern(numbers) for leads who are converted or not.
Magazine	Categorical		
Newspaper Article	Categorical		
X Education Forums	Categorical		
Newspaper	Categorical		
Digital Advertisement	Categorical		
Through Recommendations	Categorical		

Figure 1: Features in public dataset

3.3 Initial Analysis

For our initial exploratory analysis, we plotted and evaluated at all the features visually to identify those that seem most influential on the conversion rate.

We can see from Figure 2, clients get to know X Education mainly through Google, direct traffic, Olark chat and reference. Among these, Olark Chat has the worst converted rate with 25.56%, which means more efforts should be put on web chat (instant response, offer a survey for feedback, etc). On the contrary, two sources show significantly high conversion rates, Welingak website (98.44%) and reference (92.53%). This implies people are more likely to buy the courses if they are recommended by authorities or existing clients of X education.

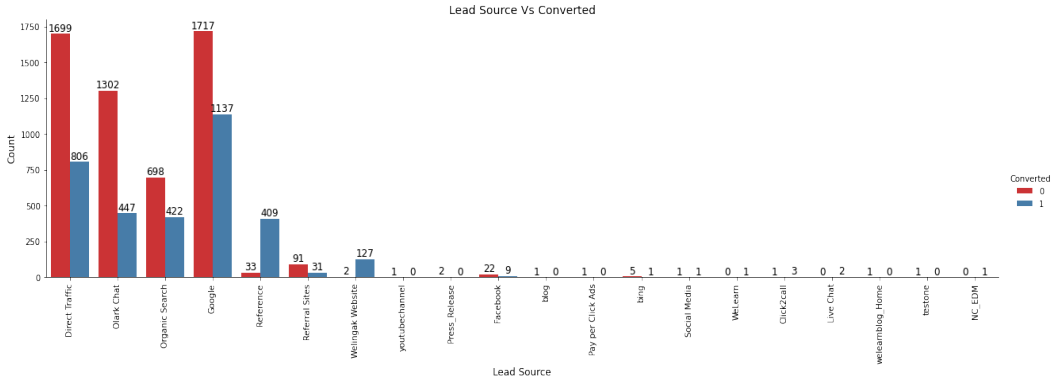

Figure 2: Lead Source

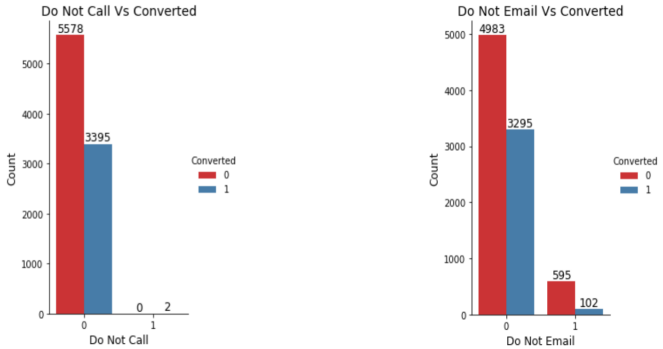

Figure 3: Do Not Call, Do Not Email

From Figure 3, the left chart tells us no one dislikes phone call, which means people prefer phone conversation instead of emails. It is noticeable that leads through phone conversation are more successfully converted. It is interesting

that almost no one refuses to receive phone calls from X education, which indicates telesales are more popular than emails.

The unemployed contribute the main sales, and there are two types customers whose converted rate above 90%, the working professional and the housewife (100%), as shown in figure 5.

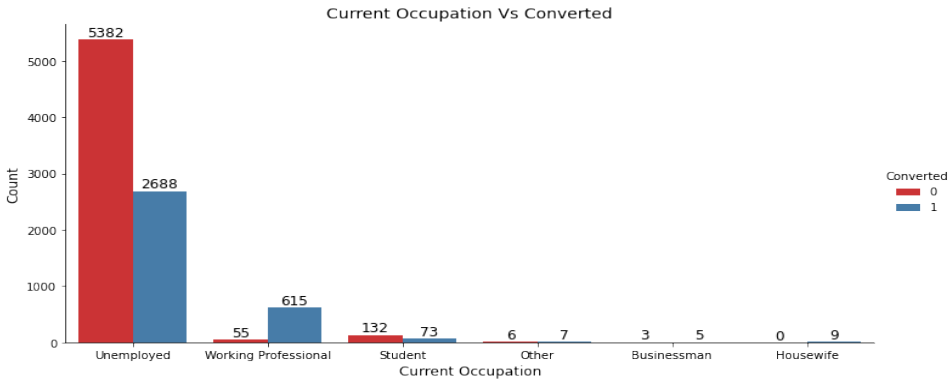

Figure 4: Current Occupation

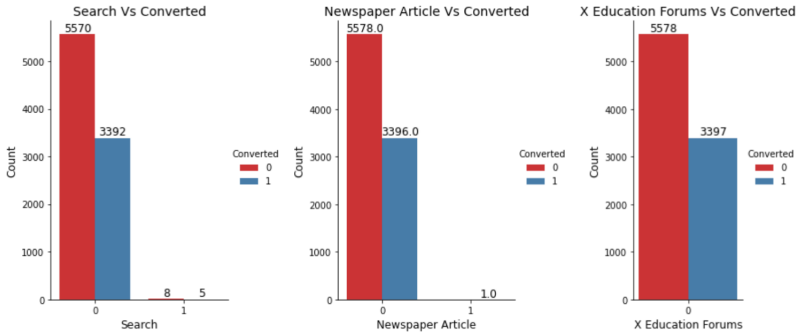

Figure 5: Search, Newspaper Article, X Education Forums

From figure 5 and 6, the factor of whether the potential customer had seen the ad in any of the listed items is not very significant. Those sources have almost the same pattern in terms of lead conversion rates. However as we can see in figure 7, people who spends more time on their website are more likely to be converted.

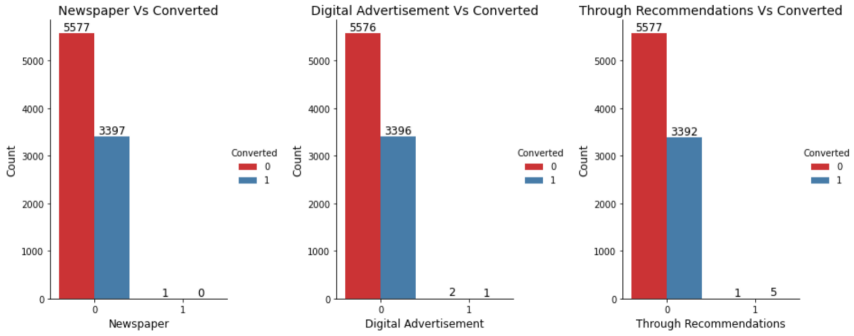

Figure 6: Newspaper, Digital Advertisement, Through Recommendation

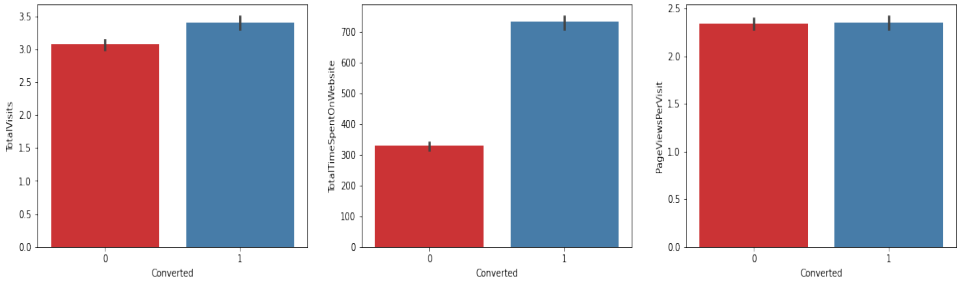

Figure 7: Total Visits, Total time spent on Website, Page Views per Visit

3.4 Missing Values

Missing data would lead to significant bias in our machine learning model. In the case of our client, many donors may not like to provide the financial information due to privacy reasons which could affect determining wealth information of the donors. As we see from figure 8, 17 features have missing values. That is a large proportion missing, if we only keep the complete instances, we could lose a lot of important information. We tried both k-means and imputation approaches to deal with the missing values.

3.4.1 Random forest & K-Means

The random forest approach treats features with missing values as response variables and uses other features without missing values to obtain the estimations of these missing values [10]. Similarly, K-Means also estimates missing values through other features without missing values. The difference is that K-Means classifies the missing data by clustering, and then calculates the mean value of the corresponding feature in the corresponding classification as an estimation of the missing value [5].

Features	# rows with missing data	Percentage of missing data
Lead Source	36	0.4%
TotalVisits	137	1.5%
Page Views Per Visit	137	1.5%
Last Activity	103	1.1%
Specialization	1438	15.6%
How did you hear about X Education	2207	23.9%
What is your current occupation	2690	29.1%
What matters most to you	2709	29.3%
Tags	3353	36.3%
Lead Quality	4767	51.6%
Lead Profile	2709	29.3%
City	1420	15.4%
Asymmetrique Activity Index	4218	45.6%
Asymmetrique Profile Index	4218	45.6%
Asymmetrique Activity Score	4218	45.6%
Asymmetrique Profile Score	4218	45.6%

Figure 8: Missing values

The results shown in table below outlines comparisons between the two mentioned approaches for imputation, as well as results from merely removing the missing values, without implementing any imputation. The data was fitted through logistic regression, which serves as a baseline in our model approach (for details see 4 Methods), the accuracy of the model is shown in the table. We can see that clearly imputing the missing values improves the performance of the model significantly, however it does depend of which scoring model is used. Similar imputation approaches and evaluations will be conducted on the actual client data.

Approach	Logistic Regression (Baseline)		
	Precision	Recall	F1socre
K-Means	0.7229	0.7865	0.7533
Random Forest	0.7201	0.78	0.7489
No Imputation	0.5528	0.7367	0.6366

Figure 9: Missing values experiment result

4 Method

The dataset was randomly split into 66% and 34% for the "Train" and "Test" datasets respectively. Then two classification algorithms, logistic regression and decision tree, were used for prediction,

4.1 Logistic Regression

Logistic regression model was used to examine how significant the features are to determine the success of conversion, and the results can be used for comparing the important features found in other models. The feature is significant if the p-value is less than 0.05.

4.2 Recursive partitioning - Decision Tree

A decision tree is a decision support tool that uses a tree-like graph of decisions and their possible consequences [8]. Recursive partitioning creates a decision tree by splitting it into sub-categories based on several divided independent variables. The process is termed recursive because each sub-category may in turn be split an indefinite number of times until the splitting process terminates after a stopping criterion is reached [15].

4.3 Results

	Logistic Regression	Decision Tree
MSE	0.1197	0.1274
Precision	0.925	0.915
Recall	0.925	0.915
F-Measure	0.925	0.914
TP Rate	0.925	0.915
FP Rate	0.090	0.110
AUC	0.972	0.931

Figure 10: Comparison between different classification algorithms

From the table above, the logistic model performed better in all aspects. To be more specific, the smaller the MSE the better the model fits. A high precision means that the algorithm returns more relevant results than irrelevant results. The Recall is the proportion of actual positives that was identified correctly. A high recall means that the algorithm returns most of the relevant results. Thus the logistic model has a higher value in all these fields, which

means that the logistic model produces more relevant results more accurately than the tree model. Moreover, an ROC curve plots TP Rate vs. FP Rated at different classification thresholds. The AUC (Area under ROC Curve) measures how well a parameter can distinguish between two diagnostic groups. The AUC of a model is close to one means it has excellent measure of separability.

5 Conclusion

So far in this project, we have performed exploratory analysis on a sample dataset, examined several predictive models and started the privacy impact assessment with the project team. We are planning to fit models with the actual datasets once we obtain the data. Our aspirational goal is to optimise sales strategy and recommend effective engagement models.

6 Timeline

Our specific roles are detailed below, please refer to figure 11 Gantt chart for more:

- Raelene: Team leader, organise meetings, tracks and guides progress. Introduction, literature reviews and analysis. Proof reading, finalises the report and presentation.
- Bing: Our privacy assessments officer. Does hypothesis and analysis, provided guidelines to all tasks, double checks, finalises report and presentation. Reference search.
- Xin: Initial data analysis and exploration, created Gantt chart, reference searching and our scrum master.
- XinHui: Handles missing data, exploration and analysis, propose initial exploration of the methods. Reference search.
- Ziyan: Final exploration of methods. Fitting models, evaluate and analyse results. Reference search.

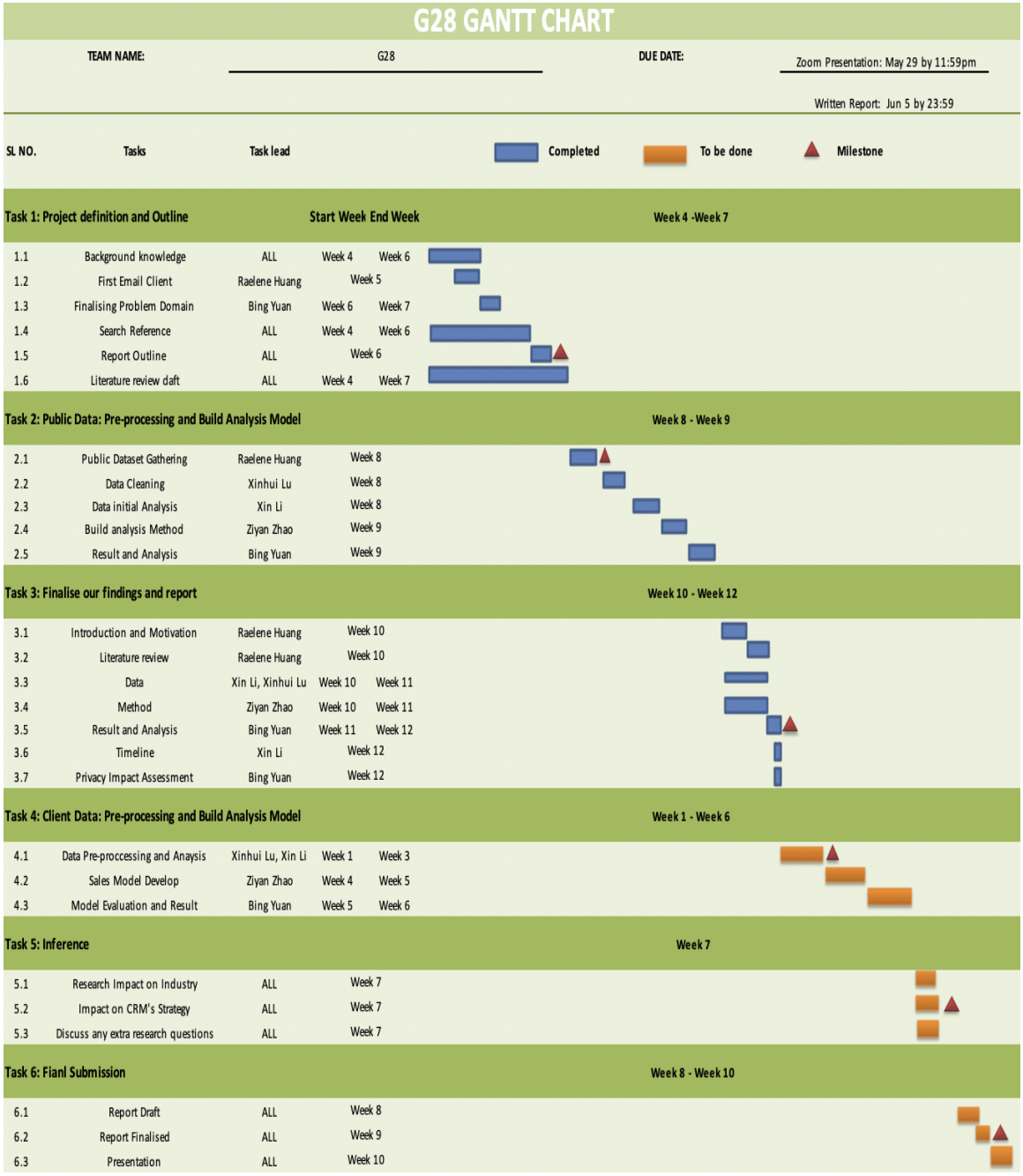

Figure 11: GANTT Chart

References

[1] Lead scoring with dynamics 365 marketing. *Dynamics 365 Marketing*.

[2] Onecrm. *University of Melbourne*.

[3] Predictive sales analytics forecasting. *Performiture*.

[4] Protecting customers' personal information. *Office of the Australian Information Commissioner*.

[5] Boyan Angelov. Working with missing data in machine learning. 2017.

[6] Ashish. Lead scoring. 2019.

[7] A. M. Baker, F. C. Hsu, and F. S. Gayzik. A method to measure predictive ability of an injury risk curve using an observation-adjusted area under the receiver operating characteristic curve. *Journal of biomechanics*, 2018.

[8] Rajesh Brid. Decision trees—a simple way to visualize a decision. *Medium*, 2018.

[9] J. V. Carter, J. Pan, S. N. Rai, and S. Galandiuk. Roc-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, 2016.

[10] W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, and D. T. Bui. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Surgery*, 2017.

[11] Sunny Chopra. How to handle missing data in machine learning: 5 techniques. 2018.

[12] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, 2001.

[13] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning. *New York, NY, USA: Springer series in statistics*, 2017.

[14] M. Kuhn and K. Johnson. Applied predictive modeling. *New York: Springer*, 2013.

[15] S. Landau and S. Barthel. Recursive partitioning.

[16] Yong S Lee. The Sustainability of University-Industry Research Collaboration: An Empirical Assessment. *The Journal of Technology Transfer*, 25(2):111–133, June 2000.

[17] Z. G. Liu, Q. Pan, J. Dezert, and A. Martin. Adaptive imputation of missing values for incomplete pattern classification. *Pattern Recognition*, 2016.

[18] I. Michiels. Lead prioritization and scoring: The path to higher conversion. *Aberdeen Group*, 2008.

[19] Robert Nygård. Ai-assisted lead scoring. *Åbo Akademi University*, 2019.

[20] J. Yan, M. Gong, C. Sun, J. Huang, and S. M. Chu. Sales pipeline win propensity prediction: A regression approach. In *2015 IFIP/IEEE International Symposium on Integrated Network Management (IM)*, pages 854–857, 2015.