CITY, UNIVERSITY OF LONDON
INM373 RESEARCH METHODS AND PROFESSIONAL ISSUES | 2022/23
Research proposal for INM363 MSc Individual Project (supervisor: Oleksandr Galkin)

Conversion: A Higher Rate and Quicker Time to Convert, Through Understanding Driving Factors.

1. Introduction

The applications of predicting Conversion Rates (CR) and Time to Convert (TTC) are important but not limited to several industries, including marketing, retail and finance. Explaining predictions is seemingly becoming just as important as creating accurate predictions in all aspects where Machine Learning exists. I want to take into consideration other variables that impact CR, but my focus is on the Product Team and predicting CR and TTC based on Product Updates.

1.1 Aims, Research Questions and Objectives

Currently there are no explainable models for predicting CR and TTC in the embedded finance discipline.

YouLend operates on a loan funding and repayment model, where merchants are the end users. This study aims to predict and understand what drives conversion and time to convert monthly as it is important to the business to identify areas of profitability as a higher accuracy of predictions would mean a larger ROI.

The research questions I propose include:

- How can we improve on the current practice of understanding and predicting CR & TTC of embedded finance platforms?
- How effective is my model at predicting conversion in comparison to the latest research in literature?

To answer these questions, the following is a list of objectives:

1. To understand the definition of conversion and gather all metrics that are suspected to contribute to conversions across the business.
2. To identify the transferability and limitations of current approaches
3. To compare different methods for gaining insight and prediction of conversion driving factors and time to convert.
4. To evaluate the extent to which the prediction models successfully address the issues identified in Objectives 2 and 3.

1.2 Outcomes, deliverables and beneficiaries

The main outcome/deliverables are:

- a new product: predicting TTC monthly.
- a re-interpretation of an existing theory: expanding on state-of-the-art prediction models.
- improved model: in comparison to the current YouLend CR model.

Additionally, another expected outcome includes addressing the hypothesis: Product Releases have a relatively low impact on CR and a high impact for TTC in comparison to other factors.

The *immediate* beneficiary is YouLend: Head of Product and merchants. The *wider* beneficiaries include other researchers that build on my research, and the embedded finance domain.

2. Critical Context

Most Literature focused on CR comes from the Retail and Marketing domain (Cezar and Ögüt 2016), and I've not come across anything suited for Embedded Finance (FinTech). Additionally, there is a lot of ML research on prediction of Loan Defaults and Credit Risk, rather than Loan Funding. There is little to no literature on predictions of TTC. TTC can also be estimated using Cox's Survival Analysis, but again there is little literature for Survival Analysis within the Embedded Finance domain. Machine Learning models have proven to be useful, but it is difficult to understand how the results were concluded (Rai, 2020). This study expands on research from the *time-series forecasting* and *marketing/finance* domains.

2.1 Conversion Rate (CR)

Cezar and Ögüt (2016) state how most previous research have investigated the number of sales rather than conversion rates. There are different ways to measure CR depending on the data available. The researchers used a window technique as browser IDs weren't tracked. This means instead of connecting a single browser to a deal/lead, the number of browsers is calculated in 10-minute intervals. Conversion funnels have been used to understand consumer behavior in past studies (HOBAN and BUCKLIN, 2015). Lee et al., (2021) research highlighted the lack of research in predicting customer conversion behaviour where most previous studies have focused on conversions using clickstream data and customer engagement metrics.

Different attributes can impact CR ranging from psychological to technical, as observed by Atulkar and Singh (2021) who studied 8 attributes on food ordering apps.

2.2 Machine Learning Models

Cezar and Ögüt (2016) studied two fractional models for conversion rate: a regression model with beta distribution and a quasi-maximum likelihood estimation model. Many papers in this area of regression modelling also used the beta distribution model which stems from Kieschnick and McCullough (2003). Cezar and Ögüt (2016) used statistical tests to ensure that the models were a good fit and stated that their results were robust as their results were consistent from both models.

The target variables in my study are both continuous and due to the nature of data my research is focused on regression models over classification. Lee et al., (2021) compares and interprets 8 different classification models for predicting online consumer conversion. The authors highlight the lack of past research for predicting conversion specifically as well as the inability to explain why the models have been suitable and interpret the results. They dealt with data imbalance, which isn't necessary for regression models. Lee et al., (2021) also makes the point of setting optimal hyperparameters for each model to compare machine learning models as performance can vary depending on the model. However, the caret package was used to compare models and

XAI (Explainable Artificial Intelligence (Lee et al., 2021)) was used to aid explainability in the context of retargeting advertisements (more on this in Section 2.3 below).

Liu et al., (2019) uses survival analysis on P2P (peer-to-peer) lending platforms. Here Liu et al., (2019) examines the probability of survival through time. The data was collected and cases where the platform did not fail were censored. The authors used three models to validate their hypothesis, including Life Table, Kaplan-Meier Estimator and Cox proportional hazards regression model. This study can be adapted to examine the probability of converting through time.

Yang et al., (2020) considered using Cox's Survival Analysis, but instead opted to use a Functional Linear Regression model to predict conversion time to Alzheimer's Disease (AD). Due to the nature of their study, their model was deemed a better fit to predict the onset time of AD.

My findings thus far look at predicting singular metrics, such as CR or TTC independently. I am yet to come across a study that builds a model(s) that can be used for both. Additionally, most of my research in the embedded finance (even finance) domain have found that there are many studies on predicting loan defaults (Xu, Lu and Xie, 2021) and little on customer conversion and TTC.

2.3 Explainability of ML models

Survival Analysis can produce the coefficients of each of the variables of the model – hence interpretable/transparent/explainable.

Lee et al., (2021) explained (using the term "interpret") their models using two XAI methodologies; XGB Importance Analysis to identify importance of variables, and SHAP (Shapely, Additive exPlanations) to interpret the output of the model. Both methodologies also incorporate visual charts to further aid analysis and explanations.

To summarise, this study will be expanding on current literature and focusing on the technical viability of predicting CR and TTC in embedded finance, as well as the ability of these models to be well-interpreted (explained). The different prediction models will also be compared and evaluated. Both focus areas demonstrate how I will be contributing to knowledge.

3. Approaches

This will be a Design and Create study, which will be broken down into five steps as suggested by Oates (2006): Awareness, Suggestion, Development, Evaluation and Conclusion. You will also find discussions on ethics, methods and limitations. Following this approach ensures my Research Questions are answered through achieving my Research Objectives.

3.1. Awareness

As with any tech company, YouLend have a product backlog in which each item can impact how many merchants are funded and how quickly this happens. As observed in the Critical Context Section, there have been many studies that can state other variables which may impact CR &

TTC. The Awareness step will allow me to achieve Objectives one and two through gathering requirements, researching and exploring the data.

3.1.1 Gathering Requirements

As my research project is sponsored by YouLend, it is important that I define my target variables correctly and refine the scope of the project accordingly. There are different CR metrics, such as: Overall (Browser to Funding Received), Step (Each stage in the Consumer Funnel), Partial (Lead to Funding Received), Initials (a new Lead) and Renewals (a returning Lead). Additionally, the product backlog items should be defined and broken down into bug fixes and releases. This initial stage will include a thorough review of all potential variables to include in the prediction models in addition to the product variables, such as a new sales process.

Moreover, I will also be assessing resource requirements in this step. For example, I will have a hardware limitation as I will be using YouLend's Windows laptop which may be computationally slow and therefore I will only be focusing on Supervised Machine Learning models for this study.

3.1.2. Research

In order to identify the transferability and limitations of current methods, I will be building on my qualitative research (see Section 2) and reviewing state-of-the-art for supervised machine learning regression models and explainability.

I will also do the same for the current YouLend model for predicting CR, as this will likely be my baseline model to compare to providing that it's of similar nature to my chosen CR metric(s).

I will be understanding current evaluation methodology for such time-series forecasting and regression models.

3.1.3 Assessing and Documenting Data

YouLend has two sources of data (CRM and platform), both of which will likely be used. However, there are challenges to overcome at the start as there is no clear mapping between these two sources and neither is there a clear mapping between "Leads" to "Loans" (funding received) which are both on the platform. I will be taking our platform data as the source of truth going forward. Data will be extracted from the CRM and platform in JSON files from June 21st, 2021, onwards (as this is deemed to be post-covid/lockdown). If any further data becomes available during my project, such as landing page or A/B testing data, I will take this into consideration (if my planned schedule allows) as this could benefit my CR and TTC predictions. It is also here that I will curate a vocabulary list for such terms as "Lending Partners" and "Merchants".

3.1.4 Ethics Review

The Ethics Review form has been appended to this proposal (see Section 7). With regards to data privacy, it should be noted that all the data I use will be hashed and it is in my nature to not share or discuss anything that may compromise the privacy or competitive advantages at YouLend. This will be confirmed before the start of the project in an agreement of the intellectual property rights and confidentiality.

3.2 Suggestion

Once most of the Awareness step is complete, I can then begin the design phase of this study and start tackling Objective 3. The Suggestion step demonstrates where I will offer a tentative idea by designing specifications based on the previous step and on-going research.

My initial key considerations are listed below.

- At the highest level, what will the inputs and outputs of the prediction models look like?
- What determines enough insight gained using model explainability?
- What limitations in current approaches can my model overcome?
- Are the models credible, robust, convincing, repeatable?
- Data considerations: (in the Development section)
- What are the suitable libraries and packages to create and evaluate the models?
- What should the confidence interval threshold look like? E.g., probability of X customer converting in the next month.

My research is comprised of quantitative research comparing and evaluating supervised machine learning models. It is limited as I am only using one data generation method, however I have the capacity for this study to include strategy triangulation by incorporating a case study.

Based on the literature I have read thus far I will choose to keep my methods similar to those in current literature, whilst keeping up to date with current literature to ensure credibility of my design. The advantage of currently choosing this method is comparing findings in literature fairly, however the disadvantage is potentially compromising results by not using a less-used research methodology. My method will also assess generalizability after the development stage.

3.3 Development

The Development step of my approach is how I will achieve Objective 3. I will be following the Agile Development Method, releasing an MVP (Minimal Viable Product) of my model and incrementally iterating to improve my study, which means increasing my awareness and refining my suggestions. I may find after data pre-processing that my suggestion needs refining and hence, I aim to communicate with my Supervisor, Oleksandr, every fortnight.

It is within this step where I will implement my proposed idea and create models to compare. This will also involve parameter tuning and explaining the models.

Development will first begin with data pre-processing. I will be verifying the data, such as checking the column types. This will be followed by feature engineering and creating (temporal) sequencing events which are assumed to contribute towards conversion, such as a new product release. Here I will consider adding cycling embeddings as Panay et al., (2021) studied in a similar task, as well as adding my target variables (CR and TTC). Before completing my data preparation, I will be using visual aids to check for any correlations in my data set.

To round of this sub-section, I will be exploring the data using more visual analysis and searching for any obvious past trends in the data that may inform or impact the CR and TTC metrics. Firstly, I will observe the distributions of my target variables, and explore the relationship between the 2 variables if any correlations are highlighted in the previous step. Next, I will use Python libraries to aid my exploration, including but not limited to MatPlotLib and SkLearn. Different variables will be mapped over time, for example observing how different Lending Partners affect CR and TTC through time.

3.4 Evaluation

Objective 4 will be addressed through an evaluation methodology as described by Oates (2006). Here I will examine the model's worth and assess any deviations from the likely results I expected. The model evaluation criteria will be selected from literature, for example using MSE (Mean Squared Error) and R2.

The model explanations will be evaluated too.

Assess generalizability. -> what can be learnt from my work that has wider applicability than just your single project.

Also, subject to supervisor's input, my project schedule will be aligned in order to collect test data while I am getting married and, on my honeymoon, to further evaluate and demonstrate the usability of my model(s).

3.5 Conclusion

The study can then be concluded by consolidating the results and any unexplained results can be deemed subject to future research.

Discuss Research Objectives: Will YouLend use the model(s)?

Can my model be extended by scope to gain further insights on what may perhaps be seen as private to YouLend and hence to report on in this study.

4. Work Plan

Milestone’s and progress meetings with supervisor + speak every 2 weeks.

start Mon 5th June. No tasks to occur during Aug 17th – Sept 3rd and during Sept 11th – Oct 7th.

Awareness (3.1) – 5th – 23rd Jun
⇒ Define Business Requirements : 5 days 5th ->9th
⇒ Source Relevant Literature : 6 days -> 6th -12th (Busy June 10th)
⇒ Identify common themes and weaknesses : 3 days -> 12th – 14th
⇒ Write Literature Review : 6 days -> 14th – 23rd
⇒ Data Collection : 3 days (+1 day slack) -> 15th – 19th (+ 20th)
MILESTONE: initial phase complete -> 20th
Supervisor Meeting : 1 day -> 20th

Suggestion (3.2) - 21st – 28th Jun
⇒ Design specifications : 2 days (+ 1 day slack) -> 21st – 26th (with slack, Busy 23rd – 25th)
=> Produce tentative idea : 2 days (+ 1 day slack) -> 26th – 28th (with slack)

Development (3.3) Jun 29th – Aug 15th
⇒ Data Preparation : 7 days -> 29th – 6th
⇒ Data Exploration : 10 days -> 4th – 14th
=> Refine tentative idea : 1 day -> 17th
MILESTONE: Design phase complete : 1 day -> 18th (Jul)
=> Write up data exploration and proposed idea : 4 days -> 13th – 5th (Aug)
=> Create proposed idea : 14 days -> 18th – 5th (change to 21 days: 12th)
-> as this includes explainability, I can start with testing models at the same time
ADD testing & validating results : 7 days -> 9th – 16th
MILESTONE: Creating phase complete : 7th
YouLend Sponsor meeting : 7th
=> write up : 6 days -> 1st – 15th

Iterate (3.1-3.3) Aug 8th -
=> Research : 4 days -> 8th – 11th
=> Refine Design Spec : 2 days -> 11th – 12th
=> Propose refined idea : 3 days -> 12th – 15th
MILESTONE: Design refinement complete -> 16th
=> Create refined idea, part 1 : 6 days -> 4th – 9th (Sept)
=> Create refined idea, part 2 : 6 days -> 12th – 20th (Oct)
MILESTONE Create refined idea : 20th
=> write up : days -> 23rd -

Evaluation (3.4) ⇒

Conclusion (3.5)

5. Risks

Risk	**Likelihood, Consequence, Impact**	**Alleviation**	**Control**
Project delays before my wedding/honeymoon (Evolving)	3, 3, 9	Plan to complete early.	Re-align on scope with my sponsor (YouLend) and supervisor.
Pressures at work increasing (Evolving)	3, 3, 9	Add slack around expected deadlines.	Utilise time-management resources and consult supervisor.
Not enough development time (Evolving)	2, 4, 8	Use known programming language and knowledge developed from the course.	Agree with research project sponsor to prioritise academic issues over technical issues – so it won't affect report write up time.
Loss of Motivation (Internal)	2, 4, 8	Very real with ADHD so plan to body double, regular check-in with supervisor, plan to complete early, work on more than one thing at a time.	Work on anything in the project that interests me to pick up motivation again or change time plan.

Difficulty in creating/interpreting models (Evolving)	2, 4, 8	Schedule slack and time to read documentation and following the latest research.	Consult supervisor and colleagues/sponsor or re-evaluate scope.
Computer Hardware Failure (Event)	2, 3, 6	Back up code and write up, have a spare laptop.	Retrieve any lost information and consult colleagues and supervisor.
Company enters administration (External)	1, 5, 5	Nothing can be done.	Keep supervisor and course administrators informed.
Illness (Event)	1, 4, 4	Keep fit and take scheduled breaks.	Inform supervisor and apply for an EC if necessary.

6. References

Atulkar, S. and Singh, A.K. (2021) ‘Role of psychological and technological attributes on customer conversion to use food ordering apps’, International journal of retail & distribution management, 49(10), pp. 1430–1446.

Cezar, A. and Ögüt, H. (2016) ‘Analyzing conversion rates in online hotel booking: The role of customer reviews, recommendations and rank order in search listings’, International journal of contemporary hospitality management, 28(2), pp. 286–304.

HOBAN, P.R. and BUCKLIN, R.E. (2015) ‘Effects of Internet Display Advertising in the Purchase Funnel: Model-Based Insights from a Randomized Field Experiment’, Journal of marketing research, 52(3), pp. 375–393.

Kieschnick, R. and McCullough, B.D. (2003) ‘Regression analysis of variates observed on (0, 1): percentages, proportions and fractions’, Statistical modelling, 3(3), pp. 193–213.

Lee, J. et al. (2021) ‘A Comparison and Interpretation of Machine Learning Algorithm for the Prediction of Online Purchase Conversion’, Journal of theoretical and applied electronic commerce research, 16(5), pp. 1472–1491.

Liu, Q. et al. (2019) ‘Survival or die: a survival analysis on peer-to-peer lending platforms in China’, Accounting and finance (Parkville), 59(S2), pp. 2105–2131.

Oates, B.J. (2006) Researching information systems and computing. London: SAGE Publications (Book, Whole).

Panay, B. et al. (2021) ‘Forecasting Key Retail Performance Indicators Using Interpretable Regression’, Sensors (Basel, Switzerland), 21(5), p. 1874.

Rai, A. (2020) ‘Explainable AI: from black box to glass box’, Journal of the Academy of Marketing Science, 48(1), pp. 137–141.

Xu, J., Lu, Z. and Xie, Y. (2021) ‘Loan default prediction of Chinese P2P market: a machine learning methodology’, Scientific reports, 11(1), pp. 18759–18759.

Yang, S.J. et al. (2020) ‘Functional linear regression model with randomly censored data: Predicting conversion time to Alzheimer ’s disease’, Computational statistics & data analysis, 150(Journal Article), p. 107009.

7. Ethics Review Form (Version 4.4, October 2015, April 2019)

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC)

https://www.city.ac.uk/about/governance/committees/cs-research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part.
The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form*.* Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk.
The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

Part A: Ethics Checklist

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try -* https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee -* http://www.scie.org.uk/research/ethics-committee/	**NO**

1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - **https://ethics.city.ac.uk/**		*Delete as appropriat e*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriat e*
3.1	Does your research involve participants who are under the age of 18?	**NO**

3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**