INFORMATICS INSTITUTE OF TECHNOLOGY

ROBERT GORDON UNIVERSITY ABERDEEN

PREDICTING FUTURE EXPENSES FROM PERSONAL BIGDATA TO MINIMIZE COST OF LIVING RISKS WITH MACHINE LEARNING FOR EMPLOYED ICT GRADUATES IN SRI LANKA

Project Proposal

Uthistra Karunagaran

SYMBOL OF SUCCESS

Contents

1. Title
2. Introduction 4
3. Area of research....4
4. Type of Research....4
5. Research question 5
6. Rationale....5
7. Objectives....6
8. Literature Survey....6
   - Gap Analysis....7
9. Methodology....7
10. Deliverables....8
11. Tools to be used....9
12. Risk identification and Management....9
13. Timeline....10
14. Student Project Ethical Review (SPER) form....11

1. Title

Predicting Future Expenses From Personal Bigdata To Minimize Cost Of Living Risks With Machine Learning For Employed ICT Graduates In Sri Lanka

2. Introduction

Today’s individuals prone to track their progress by themselves with the help of Smartphones, with apps and GPS capabilities digitally which called as “lifelogging”. Similar way, it is important to track their expense to foresee future in a big picture. In this case, people who are with high literacy rates but coming from socially stratified classes based on wealth, “white-collar” workers with bachelor degrees are indeed in-need of to adjust their earning for different reasons.

Whereas mechanisms like “Receipt-logging, Lifelogging” from “Personal Big Data” comes to play while enabling individuals to digitally capture their own expense on their own space and know the importance of their own data“ (Cathal Gurrin, 2014) which is the intention of this research to introduce it to Sri Lankan White-collar graduates. Precise reason for mentioned target audience is because graduates are already resourced with incomes and expense statements in terms of Debit/credit card payment statements, Insurance bills, Leasing payments, Loan payments and other expenses which may not sequentially tracked and used only to make the payments timely whereas if tracked correctly can overcome cost of living risks easily.

3. Area of research

Deep Learning cum Machine Learning on Personal Big Data

Reason for using Deep learning mechanisms is because expense information can be available in different formats. Individuals with e-banking capabilities has the details in pdf formats, individuals with not signed for a e-banking facility doesn’t have the details in pdfs but can capture images on the other hand, those who are not aware of any of these capability may want to add the details manually. To analyze and support various kinds of data based on 4Vs Deep Learning cum Machine Learning used in this research.

4. Type of Research

Applied research.

This study will be focused to provide a personal forecasting platform based on current expenses with Deep Learning. Which will study the pattern of current spending by Machine Learning and then forecast the future trend considering inflation rates of past years by applying Neural Network and regression models with Time series and Naïve Bayes classifiers.

5. Research question

Today’s world, since computer storage has become a more affordable and high-end gadget; technology makes us do things by ourselves. Mainly allowing humans to self-quantify based on locations, images, preferences, health, and social statuses. Considering such self-quantification facts this study will be focus towards low/middle-class graduates of Sri Lanka who are a very crucial part of the society that in-need to carefully manage their expense for the living, as well as whom can be considered as struggling to achieve the future goals based on their preference of spending and social limitations. For example, what can be done/what needs to be done if an un-employed graduate gets a job, Similarly, what can be done again or what needs to be done if an employed graduate gets an increment/good job offer. Should they keep-up with old habits, do they do savings or consider saving the income for future targets by strictly managing their spending. Considering all these facts, a better solution would be to self-aware on personal expense now and future, manage/track them accordingly and achieve the targets.

6. Rationale

The rationale of this research proposal is that firstly, to make graduates be expense-aware and mitigate cost of living risks such as for higher studies, retirement, buying a home/vehicle, starting a business in-advance and make them financially secure in their early stages. Secondly, to get to benefit by such data to the external stakeholders. Such as Insurance Companies, Banking, Telecommunications and HR department to strategize their marketing aspects. Generally, in Sri Lankan context, there are two types of borrowings can be made to achieve financial targets. First, Bank loan; Second, microcredit from money lenders. The problem here is that, as mentioned in the evidence section graduate who is middle-class / below middle-class spend most of their earning towards the day-to-day expenses, and again pay for loan/microcredit becomes a hurdle. Even so, if someone intends to borrow money from the bank they may have to find a guarantor or mortgage their properties (Ceylon, 2017) .On the other hand, paying for microcredits also doubles the expenses. (Brooks, 2018). Nevertheless, majority studies focus the influence of numeric skills on financial behavior, hence this study is to suggest an open option to achieve financial goals based on their spending (receipt-logging) and do income-driven savings to achieve the financial targets.

7. Objectives

Aim

Predict the future expenses of graduates in-advance, control expenditure and pursue for income-driven saving to mitigate cost of living risks.

Objectives

1. To develop an algorithm to recognize future monthly spending pattern of employed graduates
2. Identifying the relationship between external factors, self-control, spending behavior, inflation rates and rationalize them
3. Recommend risk mitigation options (Insurance type, Loan/Leasing amounts)
4. Develop a mobile platform to enter expense data and to forecast future expense

8. Literature Survey

The initial idea of "Receipt-logging" is derived from a phenomenon called "Lifelogging"- whereby people can digitally record their own daily lives in varying amounts of detail, for a variety of purposes" (Cathal, et al., 2014). which used in this proposal to gather varying amounts of details relating to expenses (Toshiki , et al., 2010) for predictions and recommendation. With that said, based on the main characteristics of "Big Data" volume, variety, velocity, and veracity this "Receipt-logging" emerges as a "Personal -Big Data" for the reasons of availability of different varieties information can be generated personally. Moreover, this literature review includes different areas to support the proposal topic. Such as economy base theories, rationalization theories, Regression, Behavioral variations on spending patterns and machine learning mechanisms. To begin with, Life – Cycle Hypothesis (LCH) (Albert & Franco, 1963) posits that, an individual tends to borrow when their income is low and saves when the income is high. Relating to the suggested proposal it is must in current society regardless of low income or high income to save for an unpredictable future occurrence. Similarly, Rational Choice Theory (Gary, et al., n.d.) helps to understand that human decisions are based on rational, cautious, and logical facts which is the intention of this proposal to give an insights to individual's from their own personal information to make successful personal choice. Another relatable theory is Behavioral Life-Cycle Hypothesis because this theory mentions about 3 aspects of self-control issues in saving (Svatopluk, et al., 2015) to understand the reason which aspects affects on expense choices. As below mentioned,

- The individual spends all their current financials for the consumption
- Individual invests in a variety of assets that have different levels of temptation associated with them
- Engage in savings with a framed mentality

Furthermore, Keynesian Theory points that aggregate demands will not always cater to produced supplies (lumen, n.d.) to be specific mentions about need of knowing the risks ahead and plan the options accordingly.
Considerably, to create a prediction and recommendation model Machine learning cum Deep learning for

Text Recognition is been used to analyses different expense depicted formats (receipt images) which needs a correct identification/recognition to record the expense values. hence CTPN (Connectionist Text Proposal Network) and AED (Attention-based Encoder-Decoder technique will be used (Anh Le, et al., 2019) as a learning mechanism to identify the texts in fed data. Following that, Regression to the mean – Statistical model allows natural data pattern to be looked as real-world change (Adrian, et al., 2004), Finally, Naïve Bayes Theorem– This technique helps to recognize presented variables are independent to compared to other variables (Shubham, 2018), (Jing , et al., 2011).

Gap Analysis

Almost all above-mentioned theories and implementations does not track the expense aspect in an digitalize manner from the personal bigdata. Whereas more importance given to the sensor output data, Fitbit fitness apps, body mass index data feed personally into the third-party apps and moreover input expense for a third part app only to follow-up on daily/monthly expense. Which is the intention of this proposal being to predict and recommend different options to individuals and outside stakeholders to benefit from.

9. Methodology

Below conceptualization is drawn based on different literature reviews, empirical findings, and other observations. Depicting what variables does not depend on another, which variables influence the strength of the independent and dependent variable as well as variables that depend on other variables.

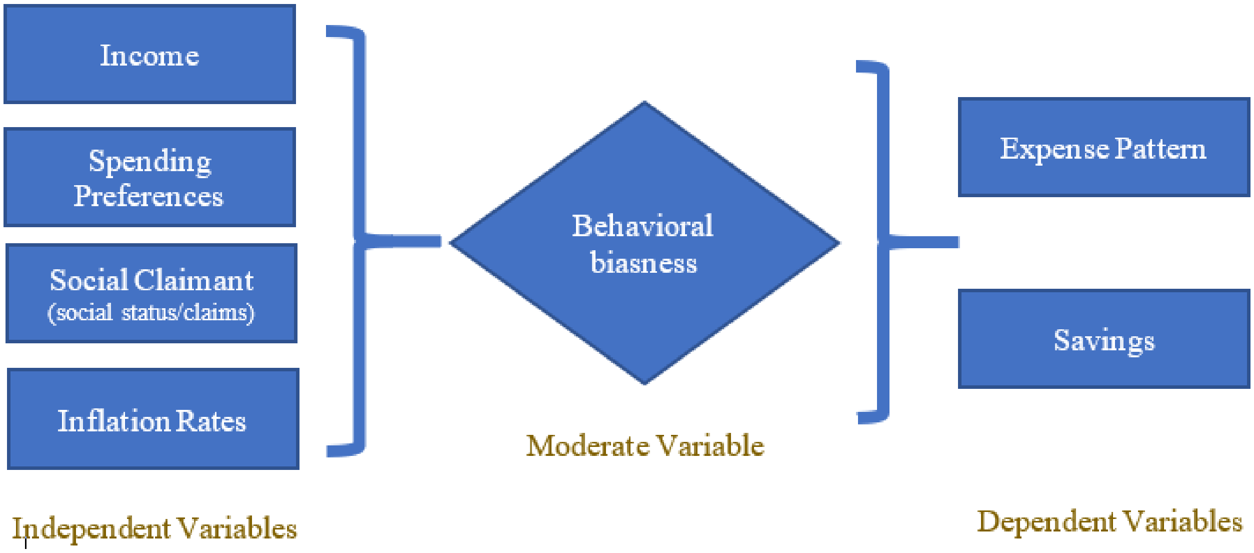


Figure 1 Conceptual Framework

Following the conceptualization, the whole process will be executed in 3 phases as the methodology.

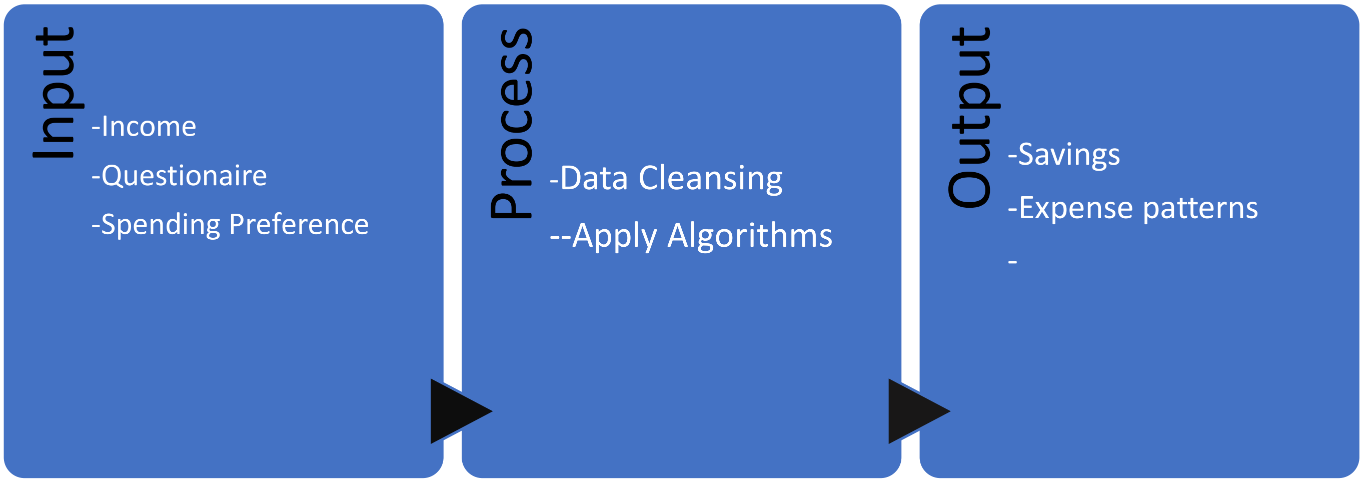


Figure 2 Phases of Execution

10. Deliverables

1. Model to identify future personal expense patterns
2. Continuous tracking of the expense categorically
3. Mobile Platform to record and visualize the patterns
4. As a by-product, individuals can act as a data broker where they can sell the personal data to a third-party by themselves

11. Tools to be used

Below depicts the tools & platform intent used to build the model. As a fact, through the empirical findings it is identified that expense information can be input or available in varied formats hence data cleaning must happen accordingly which is why weka/SQL server will be used. Also, once data is cleansed next step involves the data exploration (EDA – Exploratory Data Analysis) to understand and proceed further with machine learning tactics to identify its patterns. Nevertheless, we can incorporate the azure platform to any kind of software implementations (mobile/Web/Distributions systems) as well.
Following the EDA, algorithms (Naïve bayes),regressions ,co-relations will be predict/identify/evaluated and forecast the suggestions incorporating Rstudio & SPSS which will visualize the outputs in Power BI platform.

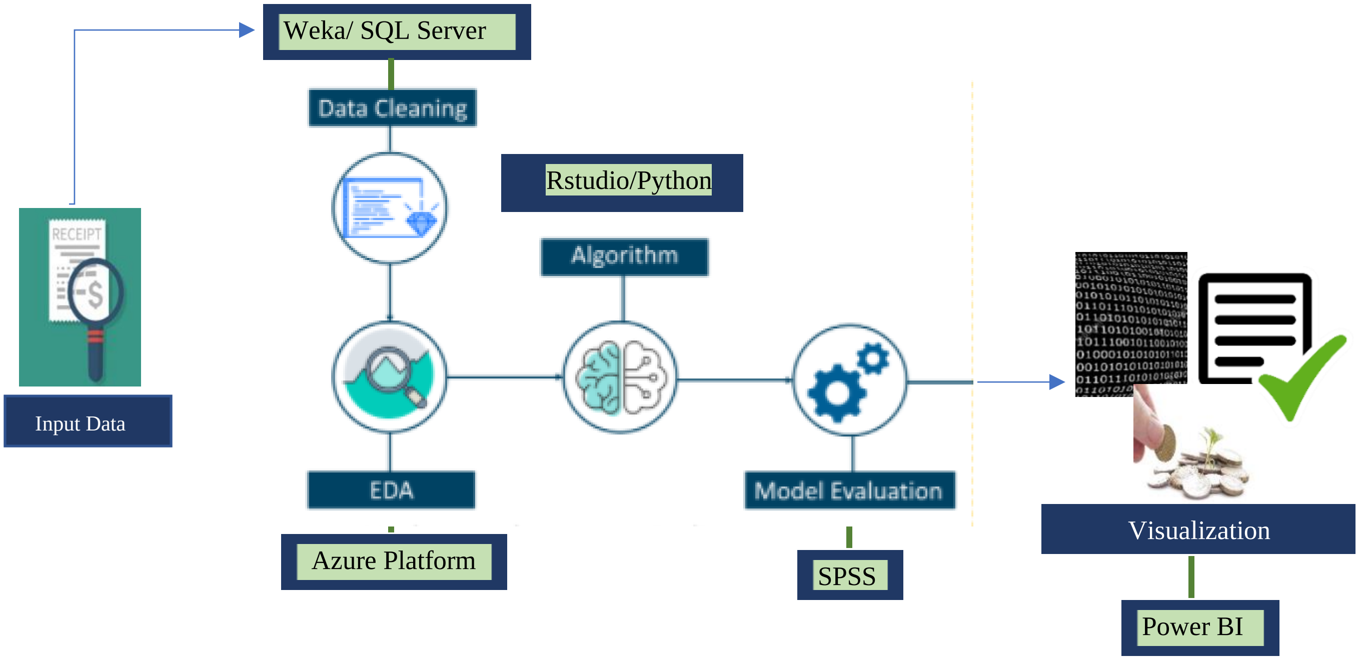
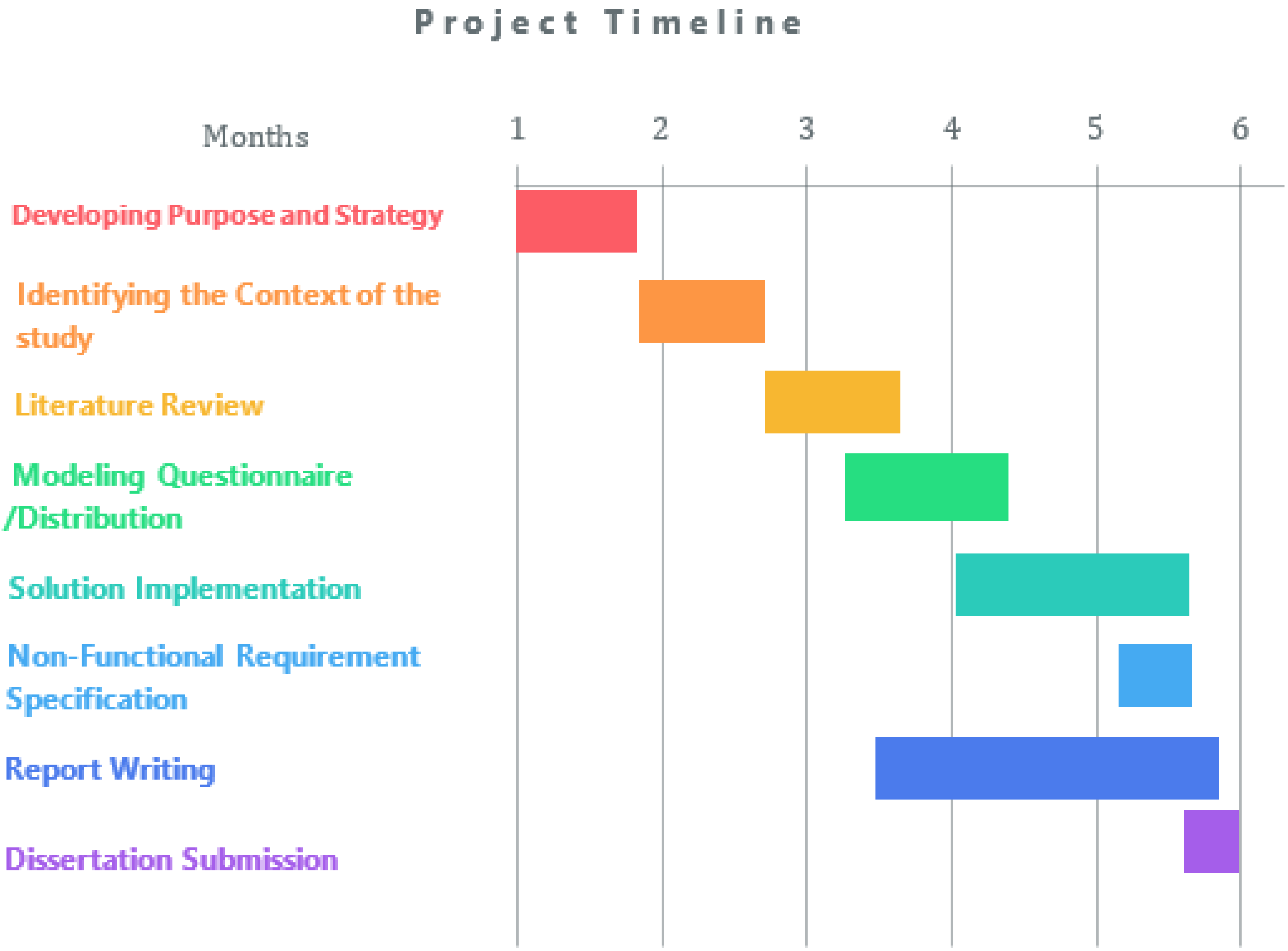


Figure 3 Tools used in Process

12. Risk identification and Management

1. Accurate Data Collection from graduates for a Time-period
2. Maintaining Anonymity of the data
3. Incorporating-Azure platform with an Android device and providing visualization platform
4. Finding outlier data and analyze its actual affect with the identified variable points
5. Identify the correct sample size

13. Timeline


Project Timeline
Months
1
2
3
4
5
6
Developing Purpose and Strategy
Identifying the Context of the study
Literature Review
Modeling Questionnaire /Distribution
Solution Implementation
Non-Functional Requirement Specification
Report Writing
Dissertation Submission