Initial Plan: Predicting driver insurance claim

Author: Seb Squire
Supervisor: Yuhua Li
Moderator: Alun D Preece
Module number: CM3203
Module title: One Semester Individual Project
Credits: 40

Contents

- Project Description – problem and motivation *Page 1*
- Main aims and objectives *Pages 1-2*
- Work plan - timeline for completing objectives *Pages 3-4*

Project Description (The problem and motivation)

A problem which drivers are likely to experience is purchasing a car, only to discover that the cost of the insurance to cover this car may be higher than expected. On many occasions drivers believe the quote for the cost of insurance they are provided with, does not reflect their driving ability and therefore is not a reasonable price to have to pay. This issue, car insurance providers can agree with as inaccurate claim predictions can raise the cost of insurance for good drivers, as well as in a lot of cases reduce the cost for bad drivers.
To further illustrate this problem, you can define some of the main factors of car insurance quotes to be determined by factors such as: years of experience, driving records and claim records. Drivers could reasonably argue that it would be unfair to have to pay a large insurance fee if they have a clean driving and claims record, especially if they have a lot of years of experience driving on roads, compared to drivers who do not.

Due to these reasons from the driver's perspective, car insurance providers are looking to specify more accurate claim predictions in order for the insurer to be able to further tailor their prices for their customers. This is in the aim that auto insurance coverage can hopefully become more accessible to more drivers. This consequently specifies the problem to be solved, where the car insurer can benefit from a machine learning model that will more accurately predict the probability that a driver will initiate an auto insurance claim in the next year. This model essentially will need to consider the main factors (as stated previously) and use probability on these factors to determine the risk involved with insuring a certain driver. In general, if a driver is indicated to be risky, the model needs to provide an accurate probability that this is true.

The main overall aims of this project can be defined by firstly, creating a model that accurately predicts the probability that a driver is likely to make a claim in the next year, and therefore this model should produce a more accurate claim prediction for a driver. Then from the basis of having this model that provides accurate claim predictions, insurance providers should be able to provide better coverage for their customers based on their personal information with regards to their car as well.

Project aims and objectives

Aims

- Generate an accurate claim prediction
    - From comparing the average statistics of given factors (e.g. Age and/or vehicle type etc.), if a driver has made a claim and the statistics proves it highly likely, the model should return true.
- Produce results to make auto insurance coverage more easily accessible to more drivers
    - If a driver has a low risk assessment for making a claim, their insurance quote should be reasonably affordable given the cost of their car.
- Make quotes fairer for drivers - From the claim prediction which is given by the predictive model, an insurer should be able to tailor the prices of auto insurance to:
    - Make it more expensive for drivers who are likely to make a claim – potentially bad drivers.
    - Make it less expensive for drivers who are unlikely to make a claim – potentially good drivers.

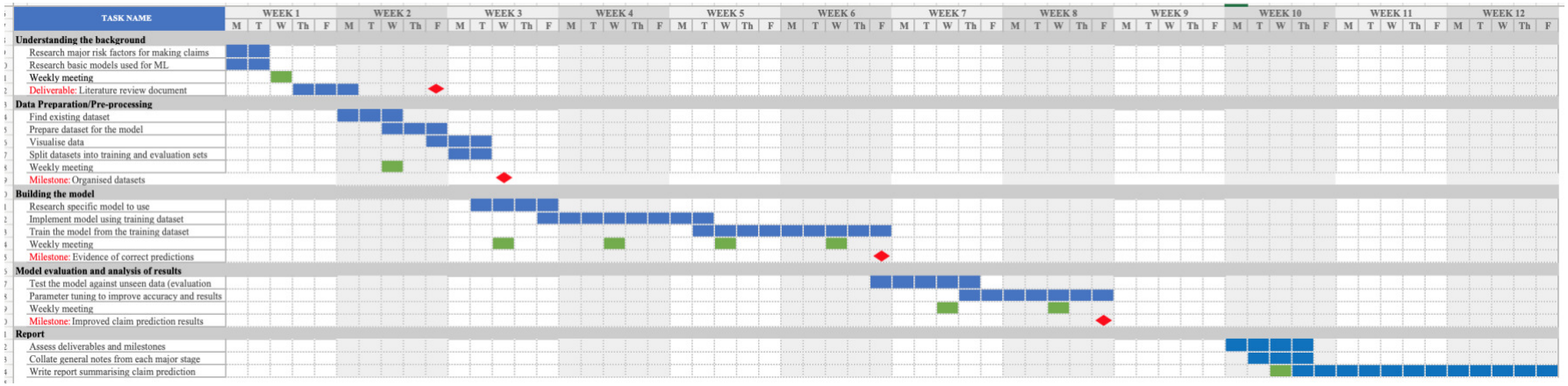
- Create a predictive model that accurately predicts the probability that a driver is likely to make a claim in the next year
    - The model should generate an accurate claim prediction, based on the training data and still provide accurate claim predictions on evaluation data as well.

<u>Objectives</u>

- Create a literature review
    - Research which specific factors could determine the risk and likeliness of a driver making a claim, as well as their existing statistics (e.g. average claim by age).
    - Research which machine-learning model is sufficient to apply the risk factors effectively for this qualitative and quantitative data.
- Collecting data and data preparation
    - Find an existing dataset (in the thousands) regarding auto insurance customers' personal information and their cars.
    - Prepare the dataset in a format of which the model will be able to interpret the data (remove duplicates, correct errors, data type conversion).
    - Visualise data to help detect relationships and provide exploratory analysis – either in the form of a table or potentially a graph database.
    - Split the dataset into training and evaluation sets.
- Design and train the model
    - Design the model - research which model could apply best to the dataset that will be used, for e.g. decision trees or using an extension to build random forests.
    - Implement the model using the training dataset.
    - Train the model based on the training dataset for as many correct predictions as possible – based on auto insurance customers who have made a claim and their personal information.
- Model evaluation & results analysis
    - Testing the model – make predictions by testing the model against unseen data (evaluation set) and compare the results.
    - Parameter tuning of the model to try and improve performance – provide more accurate claim predictions after training and evaluation on the datasets are applied.
- Report
    - Assess all the deliverables and milestones – analyse how I progressed from each stage and conclude if the results were developing for more accurate claims after each testing stage.
    - Collate these notes from major stages and expand on them for more detailed descriptions of how the model developed over time.
    - Then write a report summarising how all the stages/objectives came together to make the solution functional.

Work plan

The first deliverable that will be produced is the literature review document. This is important as it will include an overall background on the topic of machine learning, as well as the basic main factors of car insurance claims. After assessing this document, it will allow me to create a model based on these facts.
The first milestone will be the datasets organised into two main types: the training set and the evaluation set. This is important as you need to keep the two separate, in order for the predictive machine learning model to be able to predict a claim and not be biased.
The next milestone needs to include evidence of correct predictions; this will determine whether or not the predictive model is functioning correctly.
Then the last milestone is based on using the evaluation set (unseen data) rather than the training set and testing the model to see if it can generate risk predictions on the unseen data accurately. After parameter tuning the model should be more accurate and this is a major milestone which indicates if the solution works fully.


TASK NAME
WEEK 1
WEEK 2
WEEK 3
WEEK 4
WEEK 5
WEEK 6
WEEK 7
WEEK 8
WEEK 9
WEEK 10
WEEK 11
WEEK 12
M T W Th F
Understanding the background
Research major risk factors for making claims
Research basic models used for ML
Weekly meeting
Deliverable: Literature review document
Data Preparation/Pre-processing
Find existing dataset
Prepare dataset for the model
Visualise data
Split datasets into training and evaluation sets
Weekly meeting
Milestone: Organised datasets
Building the model
Research specific model to use
Implement model using training dataset
Train the model from the training dataset
Weekly meeting
Milestone: Evidence of correct predictions
Model evaluation and analysis of results
Test the model against unseen data (evaluation
Parameter tuning to improve accuracy and results
Weekly meeting
Milestone: Improved claim prediction results
Report
Assess deliverables and milestones
Collate general notes from each major stage
Write report summarising claim prediction