

For this problem using classification can be more appropriate to decide if user will file a claim. For example, through knowing user's age, health, habits and destination they will go to, we can make predictions with the provided label by creator of the dataset or teacher (Citi, L, 2020). To classify our user, we can use two different classes and for this problem classes would be "True" for users who are more likely to make claim and "False" for users who are less likely to make a claim. We have a dataset which each item have attribute named "Number of Travels per Year" and class which is either "True" and "False".

We have a simple classifier create a simple rule such as : "if $\theta > x$ then it is True else if $\theta < x$ it will be classify as False" (Citi, L, 2020).

Possible Features:

One of the reasons a customer might issue a travel claim might be health reason. In UK government's website tells health and medical emergencies should be covered by travel insurance policy such as emergency treatment and hospital bills. Therefore, it can be helpful to health information of that traveller such as if the traveller has prior health condition, current health condition and age. If the traveller more prone to have health condition while travelling, there might more chance for travel claim due to health reasons (GOV.UK,2021).

It might be important the date of the travel happens due to either security or health reasons. In 2020 due to Corona which is a disease with high rate of infection travellers are more vulnerable to being infected by Corona disease and it is possible to traveller might issue a claim. In 2016, due to terror attack insurance website named "InsureMyTrip.com" got 15% jump in calls from travellers to get insurance (Sheedy, R,L., 2016).

Risk level of the place will be visit might be important. For certain areas with such low security insurance companies refuses to make insurance (GOV.UK,2021). Last reason is the possible activities might be an indicator if traveller might issue a claim (GOV.UK,2021).

For this application I would rather choose Decision Trees. When it comes to insurance applications such as “Predict Customer Loss”, “Insurance Fraud”, “Insurance Claims” it is observed that tree-based algorithms are used alongside with SVM, KNN and Logistic Regression (Baudry, M., & Robert, C. Y., 2019).

Decision tree is a tree algorithm using divide and conquer technique and like many tree algorithms it works recursively. There will be root of the tree containing the most relationship and then it will be divided and have two node containing data connected to root. For each node this continue. The performance cost of the logarithm of number of elements within the tree, it is possible to visualize the tree and it is possible to extract simpler rules that able to generalize the problem ("sklearn.tree.DecisionTreeClassifier — scikit-learn 0.24.0 documentation", 2021).

However, there are disadvantages like it is possible to make the tree overfit and have poor predictions for future data. That can be preventable by either using pruning (Citi & Scott, 2020).

Before evaluation there should be two things be done. First for each different machine learning algorithm to split the data into test and training data. Second thing is that to make cross validation. Evaluation will be done by different metrics. These metrics will be precision, recall and f1-score. Recall is a metric if our model able to separate different classes from each other, precision is a metric if our model able to gather data which is related to a specific class and F1-score represent balance between recall and precision (Luca, C, 2020). This is significant since there might be a trade off between recall and precision. For example, a model can have one example and classified it as positive and the prediction is positive. In this case we have 100% precision but 0% percent recall. This indicate that there is a trade-off between recall and precision (Luca, C, 2020).

- 1) Baudry, M., & Robert, C. Y. (2019). A machine learning approach for individual claims reserving in insurance. *Applied Stochastic Models in Business & Industry*, 35(5), 1127–1155. <https://0-doi-org.serlib0.essex.ac.uk/10.1002/asmb.2455>
- 2) Citi, L., & Scott, P. (2020). *DECISION TREE INDUCTION* [Ebook] (pp. 33-34-35-37). Retrieved from https://moodle.essex.ac.uk/pluginfile.php/281546/mod_resource/content/0/Decision_Trees1.pdf
- 3) Citi, L. (2020) *Introduction to Machine Learning and Data Mining* [Ebook] (pp. 13-14-15). Retrieved from https://moodle.essex.ac.uk/pluginfile.php/1386526/mod_resource/content/1/CE802_Lec_Intro ML_handouts.pdf
- 4) Citi, L. (2020). *Evaluating learning methods* [Ebook] (pp. 24-25-26-27). Retrieved from https://moodle.essex.ac.uk/pluginfile.php/1404664/mod_resource/content/2/CE802_Lec_Eval_handouts.pdf
- 5) Foreign travel insurance. (2021). Retrieved 17 January 2021, from <https://www.gov.uk/guidance/foreign-travel-insurance>
- 6) Leggat, P. A., & Leggat, F. W. (2002). Travel insurance claims made by travelers from Australia. *Journal of Travel Medicine*, 9(2), 59–65. <https://0-doi-org.serlib0.essex.ac.uk/10.2310/7060.2002.21444>
- 7) SHEEDY, R. L. (2016). Insuring a Trip In Uncertain Times. *Kiplinger's Retirement Report*, 23(5), 18.
- 8) sklearn.tree.DecisionTreeClassifier — scikit-learn 0.24.0 documentation. (2021). Retrieved 17 January 2021, from <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

