# Health Insurance Cross-Sell Prediction: Predicting Whether or Not a Health Insurance Policy-Holder Will Buy Auto Insurance

Aiman Chughtai

## Abstract

The goal of this project was to create classification model to predict whether or not a current health insurance policy-owner for an Insurance company will also be interested in buying auto insurance from that company. By targeting customers who might be interested in auto insurance, the company will be able to increase revenue, build brand loyalty and understand what factors contribute to cross-selling a new service. Additionally, the company may be able to convince those interested customers to finally "bite the bullet" by offering them special promotions. The dataset I used for this project was found on Kaggle.com (https://www.kaggle.com/anmolkumar/health-insurance-cross-sell-prediction?select=train.csv) .

### Data

The dataset contains 381109 rows of customers with 10 features for each. The target variable was Response (Interested: Yes=1, No=0), which marked whether or not a customer expressed interest in buying auto insurance. The biggest issue with this dataset is that it was incredibly imbalanced; the positive class was about ⅛ the size of the negative class,

## Tools and Algorithms

- Python
- Numpy
- Scikit-learn
- Seaborn
- Matplotlib

## Results/Design: Modeling and evaluation

For this classification task, I was interested in a model that offered high predictive power but also was interpretable and could handle both categorical and continuous features.

Logistic regression was an obvious choice because it is very interpretable. As for predictability, I was interested in utilizing tree based and gradient boosting models. I therefore used a decision tree classifier, a random forest classifier, and an XGBoost classifier. For my evaluation metrics, I was interested in minimizing the amount of customers we would incorrectly predict as negative, because that would mean we would lose the potential to target a customer that actually was interested in buying auto insurance. For that reason I chose F2-score as my evaluation metric in order to emphasize recall. Each model was fit using SciKit-Learn, class imbalance was mitigated by applying class weights to the minority class, and hyper parameters were tuned for the top two models (Logistic regression and XGBoost. Ultimately hyper parameter tuning did not improve my models' performance so my final model that I chose was the Logistic regression model, with class weight "balanced", which gave me an F2 score of .617. Model was also evaluated with a confusion matrix for visualization purposes.

## Communication

A presentation/ slides,