# Springboard Capstone: BNP Paribas Cardif

Overview by Monisha Gopal

# Outline

- Background
- Data Challenges
  - Handling Data Challenges
- Data Exploration
- Machine Learning Models
- Conclusion

# Background

BNP Paribas Cardif is an global insurance company that specializes in personal insurance.

Their clients want their claims to be processed faster. However, currently claims have to go through many checks before being acceptable for payment.

So BNP Paribas Cardif hopes to use data science to determine off the bat which claims can go through an accelerated process and which can’t.

# Data Challenges

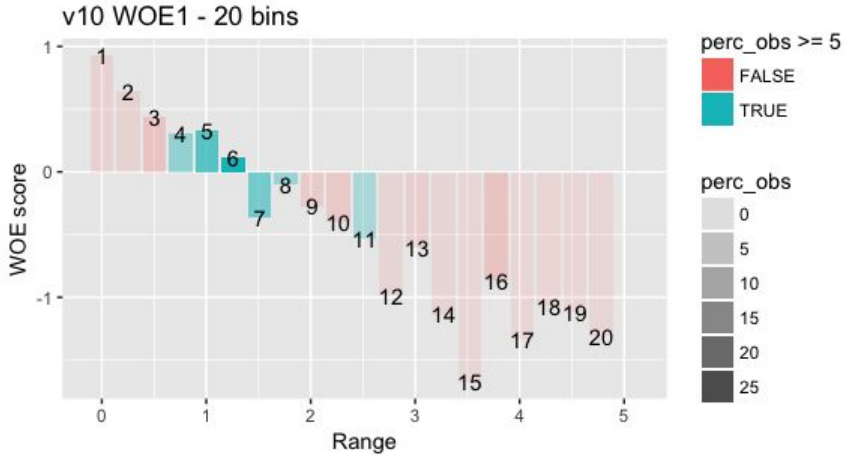
- Anonymized data
  - Don’t know what features stand for
  - Don’t know what distributions to expect
- Lots of missing values
  - Because we don’t know what features stand for, we can’t conclude the reason for missing values
  - Not representative of the population

# Data Challenges: Handling

- Anonymized data
  - Can look at values for certain features and guess what that feature stands for (i.e. Location, Claim class, etc)
  - However, can still capture trends using Weight of Evidence
- Lots of missing values
  - Remove the missing values if there are too many (more than 25%)
  - Replace missing values in continuous variables with the mean
  - Replace missing values in categorical variables with the mode

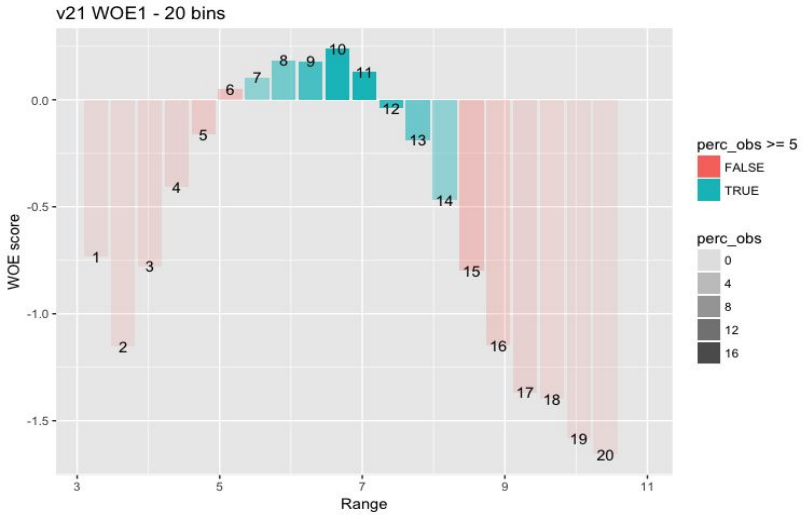
# Data Exploration

Used Weight of Evidence to capture trends




Example:
For v10, we see that if an observation has a high v10 value, it is more likely that it's *not* a claim that can be accelerated.

# Data Exploration




Example:
For v21, we can see that observations that have a v21 value between 5.25 and 7.1 have a higher chance of being a claim that *can* be accelerated.

# Machine Learning - Logistic Regression Models

Model 1 - Top 5 most important variables

- Variables:
  - v50, v129, v47, v10, v62
- Accuracy:
  - 0.6354

Model 2 - Top 8 most important variables

- Variables:
  - v50, v129, v47, v10, v62, v66, v21, v12
- Accuracy:
  - 0.6398

Model 3 - Top 10 most important variables

- Variables:
  - v50, v129, v47, v10, v62, v66, v21, v12, v72, v114
- Accuracy:
  - 0.6405

# Machine Learning: Kaggle Outcomes

| Submission and Description | Private Score | Public Score |
| --- | --- | --- |
| **model3.csv**<br>5 minutes ago by Monisha Gopal<br>Top 10 variables | 0.50107 | 0.50170 |
| **model2.csv**<br>6 minutes ago by Monisha Gopal<br>Top 8 variables | 0.50112 | 0.50169 |
| **model1.csv**<br>6 minutes ago by Monisha Gopal<br>Top 5 variables | 0.50927 | 0.50998 |

# Conclusion

Of the three models used to solve this problem, model 2 had the best results.

Even though model 3 was created using more features, the accuracy of the model wasn't significantly better than model 2's. Model 2 also did better than model 3 when predicting Kaggle's general dataset.

# Thanks