**Executive Summary**

In this project, 5 classification models were created to identify users that might stop using a service relying on personal data use. Reasons for churning were examined by performing exploratory data analysis and recommendations for improving customer relationship and user retention were made. The analysis was not specific to any individual company.   
The project draws attention to the fact that about 36% of respondents have been victims of some sort of personal data misuse. Further investigation showed people do not believe there is much benefit from the personal data use. Most of the people do not understand the current laws of data protection, and do not read companies’ privacy policies yet express interest about the current state of personal data use and regulations. These factors seem to be among the possible reasons for users’ mistrust and churn. Majority of people want more government regulation and better tools to control the use of personal information themselves. The evaluation of 5 classification models showed that Neural Network Classifier was the best candidate to address the challenge of predicting customers willing to depart due to the personal data use. The 72% recall of the positive class would allow for correct churn classification of more than 2/3 of users.

Recommendations:  
People identified as the ones likely to stop using the service could be sent informational messages with simplified descriptions of personal data use protection measures. After using marketing means to address the issue uplift model could be used to further improve the results of churn predictions. Larger sample size and additional features added by companies could improve the predictions.

**Introduction/Background of the Problem**

2018 is known as a very important year in the data sector. The Security breaches in the companies like Facebook, Quora, Starwood and etc. as well as GDPR coming into effect as a need for data use regulation attracted even more attention to the way companies work with personal data and caused the rise of distrust and uncertainty among the population. The amount of personal data used increases every year. The availability of technology, especially mobile devices and cloud computing allow for more data being created and shared. Companies with mature ethical policies abiding by the regulations would benefit from a trustful partnership with the users. Trust plays a very import role in user retention and lack of it is one of the main reasons why a client could leave the partnership. Taking steps to disclose the practices and explain honest intentions, inform the customer, possibly create regular data accountability reports to share with the clients at risk of departure could prevent their churn. Identifying who and why someone is more likely to stop receiving services because of personal data use could have practical benefits for both the companies such as stronger partnership creation, higher revenue, staying ahead of competition; and customers by receiving more personalized and up-to-date services based on the authentic information.

**Methods**

|  |  |
| --- | --- |
| CRISP-DM STAGE | METHODS |
| Business understanding | defining objectives, finding data, target variable, measures of success for the model, -ideas for models, planning |
| Data understanding | summary report for variables; categorical variable assessment; data visualizations (bar plots, pie charts); pivot tables; missing values identification |
| Data preparation | data quality analysis (accuracy, relevancy, completeness, timeliness, consistency); removing missing data; feature selection; nominal variables transformation; data partitioning; data standardization |
| modeling | 5 classifiers built, Hyperparameters tuning |
| evaluation | Results analyzed and compared with the focus on recall; Best model selection; Model’s issues fixed |
| Recommendations for deployment | Limitations and precautions identified; Consistency of the process stressed; Possible further analysis proposed |

**Results**

**for reproducibility:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| OS | Languages | Libraries | Data | Seed |
| Windows Home 10,64-bit operating system | Python(3.7.2.), R(3.6.2) | pandas, matplotlib, numpy, scikit-learn, tensorflow, keras, ggplot | [American Trends Panel Wave 49](https://www.pewresearch.org/internet/dataset/american-trends-panel-wave-49) | 4 |

**Business understanding:**

The *objective* of the project: identifying users that might stop using a service due to personal data use with the help of American Trends Panel Wave 49 dataset.

The *recall* of positive class of classification models such as Logistic Regression, Random Forests, Extreme Gradient Boosting, Support Vector Clustering and Neural Network (Keras) was chosen as a *metric* for evaluation because of the primary interest in identifying users that won’t use the service.

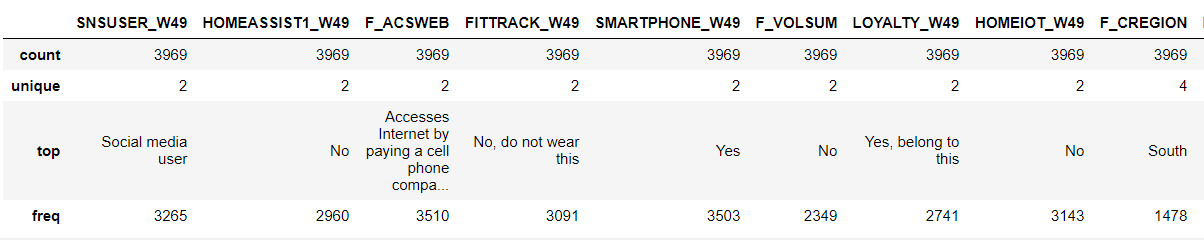
**Data understanding and data audit**

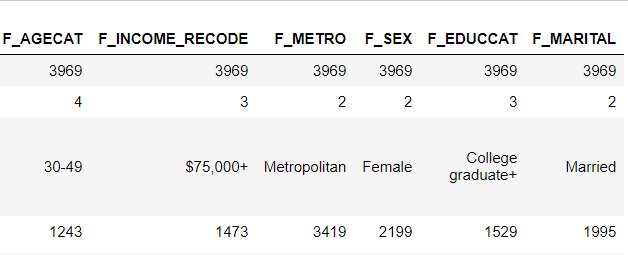
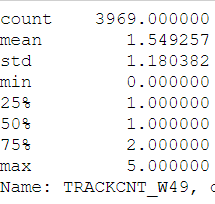
*Quality:* Data consisting of 143 variables and 4272 observations is of *high quality* without any pattern of satisficing. The sample is weighted (variable WEIGHT\_W49).

*Missing values:* A third of data has more than 50 % *missing* *values* (some questions were asked only of the respondents that answered “yes” to the previous uestion). The value “Refused” is treated as a missing value.

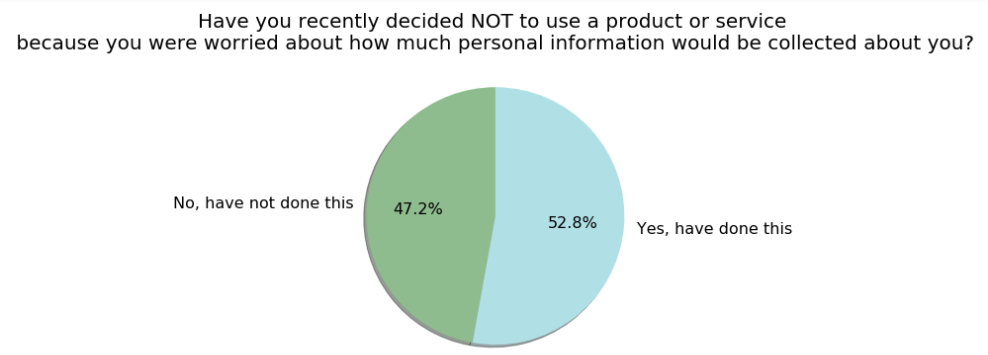
*Numeric variables*: QKEY – customer ID, TRACKCNT\_W49 and TOTALKNOW\_W49 – number of questions answered yes or correctly, and WEIGHT\_W49 – weights applied to the sample.

*Categorical* data check: absence of high-cardinality; variables with only 1 unique value are not present. Attached is a shorter version of the summary report of only the features used for modeling.



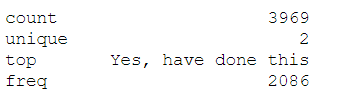
Analysis of the *Target* variable SHARE1\_W49: the balance of 2 classes visualized below:

The **classes** are pretty balanced with a slight

difference

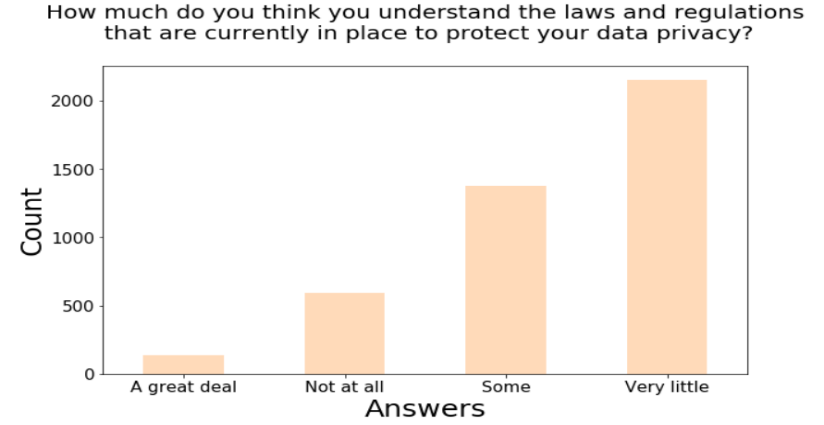
in favor of

the “Yes”

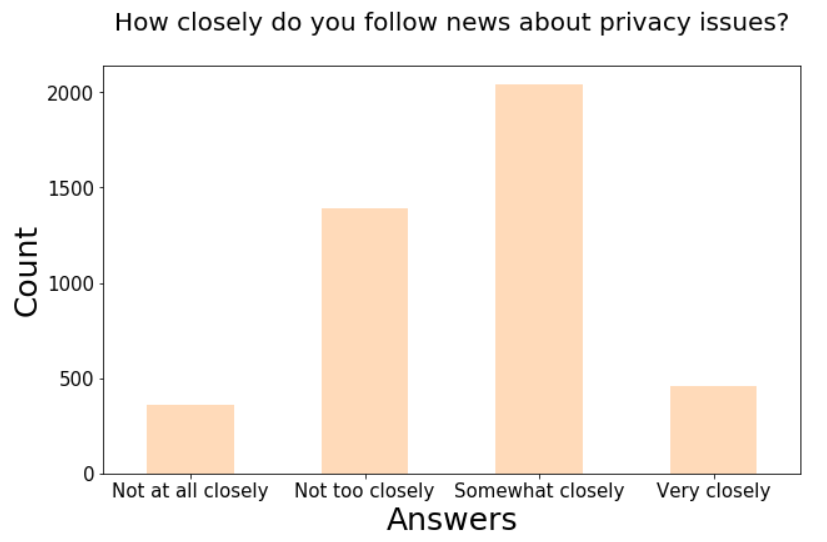


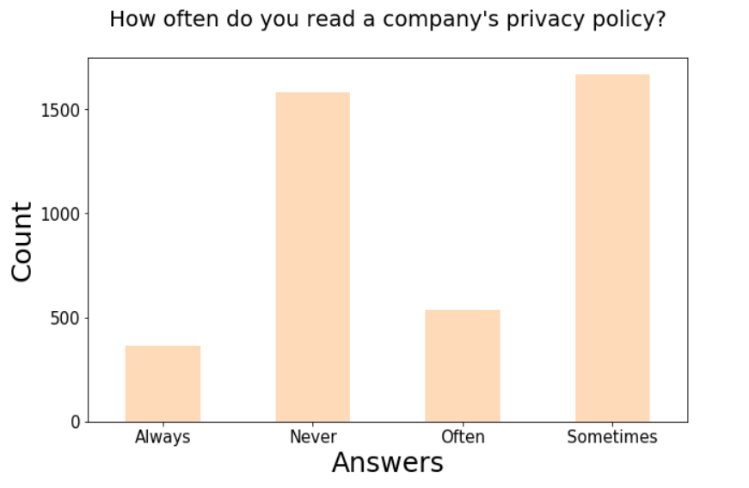
**Summary Statistics of the target variable**

**Exploring patterns and insights with visualizations:**

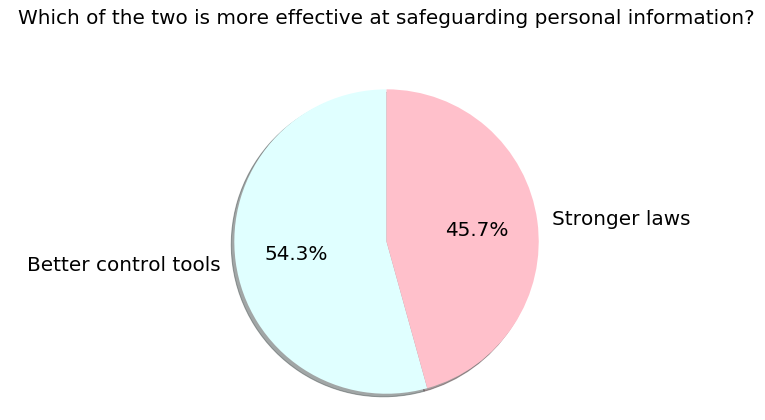


**Plot 1.** Most of the respondents have very little understanding of the laws and regulations that are currently in place to protect data privacy

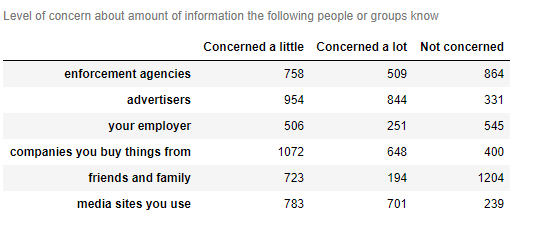
**Plot 2.** Most of the respondents follow the news about data privacy somewhat closely

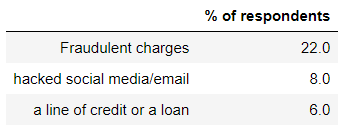


**Plot 3.** Most of the respondents only sometimes read the privacy policy and a lot of the respondents do not read it at all.



**Plot 4.** Better tools for allowing people to control their personal information themselves is a more popular choice among ways of safeguarding personal information.

 **Table 1.**The highest concern is about how much information advertisers know, followed by media sites you use and companies you buy things from.

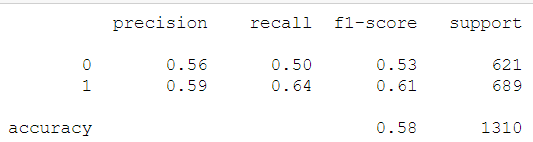
**Table 2.** 22% of respondents experienced fraudulent charges. Overall, 36% of respondents have been affected by some kind of data misuse

**Data Preparation:**

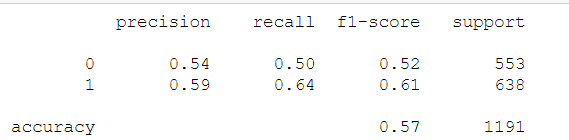
Missing values were removed.

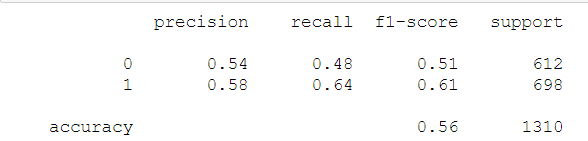
*Feature selection:* Selection of features was based on how measurable they are and whether all participants were surveyed (summary statistics and list of features could be found above).

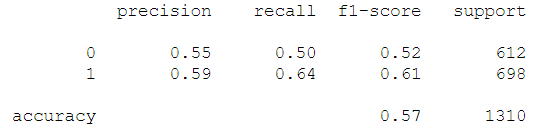
“Label encoding” was used for the target variable and one-hot encoding for the features. Features were standardized with StandardScaler()

**Modeling and Evaluation:**

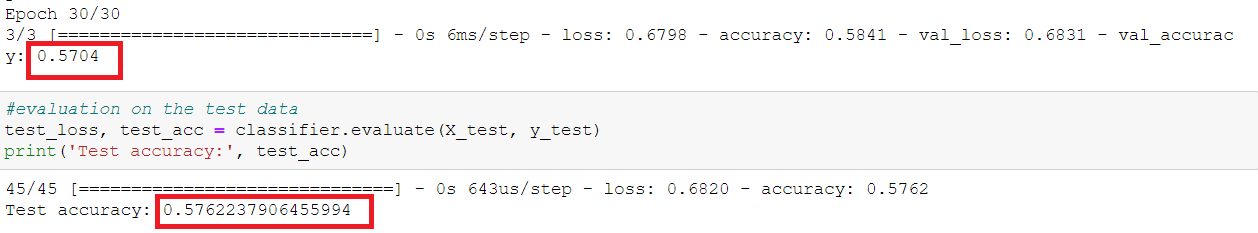
**Model 1**. **Extreme Gradient Boosting** after hyperparameter tuning with GridSearchCV. The accuracy is not high. What is more important, the recall is at 64%

**Model 2.** **Random Forests** after hyperparameter tuning (GridSearchCV). The accuracy is not high. What is more important, the recall is at 64%

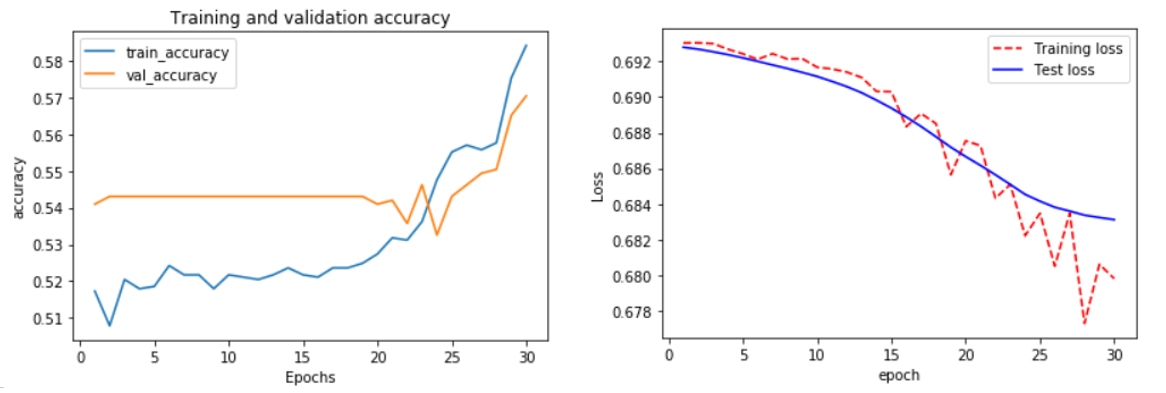
**Model 3.** Logistic Regression after hyperparameter tuning (GridSearchCV).The accuracy is not high. What is more important, the recall is at 64%

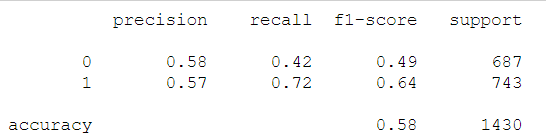
**Model 4.** Support Vector Clustering Algorithm after hyperparameter tuning (GridSearchCV). The recall is at 64%.

**Model 5.** **Neural Network Classifier with Keras** (regularization l2 and dropout of 0.55 applied)

 Accuracies of validation and testing sets are very similar

Plots of accuracies and loss showing the model is not overfitting with the training accuracy growing and being higher after some training and the loss degrading for both test and training set and the training one being smaller than the test one.



The recall for this model is the highest one out of all 5 models.

**Confusion matrix.** 536 (72%, more than two third) out of 743 users that have churned have been identified correctly. This could be very useful.

**Recommendations for deployment:**

* Model is especially useful for companies using loyalty programs and devices using personal data
* Ensuring reproducibility relying on information provided in the report is important
* More data could be collected for higher performance
* After using marketing means to address the issue uplift model could be used.

**Discussion/Conclusion**

The best model used to predict customers that stop using a service due to personal data use is Neural Network Classifier with the accuracy of 0.58. The recall for the positive class is at 0.72, which means that the model classified more than two thirds of the users that churned correctly, which is better than random guess and could be useful.

**EDA insights that could help companies understand reasons for churn and establish trust:**

* About 36% of respondents have been victims of some sort of personal data misuse.
* People do not believe there is much benefit from the personal data use yet they acknowledge that it is not possible to go about daily life without personal data use.
* Majority of people want more government regulation over the personal information use.
* They believe tools that would let them control personal data use themselves are more effective
* Majority people do not understand the current laws of data protection, yet they express interest about the current state of personal data use and regulations.
* Most of the people do not read companies’ privacy policies.
* The highest concern is about the amount of information advertising companies know;
* People are least comfortable when sharing their data with outside groups;

**Possible suggestions for marketing approach to improve user-provider relationship:**

* People identified as the ones likely to stop using the service could be sent informational messages with simplified descriptions of personal data use protection measures.
* Easy to understand information should be available for the general public to increase public awareness of various privacy policies and safety measures used by a company.

**Limitations of the project:**

* Many potential predictors were not used as they would be hard to measure by businesses
* Interpretability of the neural network algorithm is challenging.

**Acknowledgements**

I would like to express my deepest gratitude to everyone who provided me with the possibility to complete this project. I especially want to thank Professor Iranitalab for reaching out, answering questions and providing feedback.  I have to appreciate the advice and suggestions of Ambrose Cordero, Alan Danque and Rene Solis . I would also like to thank my family for supporting me and helping me meet the deadlines.

References

Pew Research Center. Data. American Trends Panel Wave 49. <https://www.pewresearch.org/internet/dataset/american-trends-panel-wave-49/>

Pandas.pydata.org. Reshaping and Pivot Tables. Retrieved from <https://pandas.pydata.org/pandas-docs/version/0.24.2/user_guide/reshaping.html>

Chambliss, Ch. Cleaning, Visualizing and Analyzing Survey Data in Python. Retrieved from <https://towardsdatascience.com/cleaning-analyzing-and-visualizing-survey-data-in-python-42747a13c713>

Droste, B. 2019. How to Analyze survey data with Python. <https://towardsdatascience.com/how-to-analyze-survey-data-with-python-84eff9cc9568>

Rectenwald, Ch. 2019. Predicting Customer Churn. Retrieved from <https://medium.com/analytics-vidhya/predicting-customer-churn-58faa29bc836>