Statistical Data Analysis - Predicting Customer Complaint

Benjamin Abrahamsen Hagen

ABSTRACT

This data analysis presents a comprehensive analysis of a banking dataset to gain insights into customer complaints and factors influencing them. It involves exploratory data analysis, and predictive modeling, with the primary aim to predict the likelihood of a customer filing a complaint based on their profile. Key findings from this analysis include the identification of groups more likely to complain, the non-significance of credit scores and satisfaction scores in predicting complaints or customer exit, and the balanced nature of the bank's reward system across different card types. A logistic regression model was developed with an accuracy of 99.95% in predicting customer complaints. The model's usefulness was further demonstrated by predicting complaints for a new customer sample.

1 INTRODUCTION

In the modern bank industry, customer satisfaction is critical to success. Understanding the factors contributing to customer complaints can help banks improve their services and stop customers from exiting the bank. This analysis is aimed at identifying such factors and developing a predictive model for customer complaints. The goals include identifying the groups more willing to complain, look into the impact of credit scores and satisfaction scores on complaints and customer exit, analyzing the fairness of the bank's reward system, and predicting future complaints. This analysis is of great importance as it can help the bank address potential issues, improve its service, and thereby increase customer satisfaction in a highly competitive banking industry.

2 DATASET OVERVIEW

The dataset used for this analysis consists of banking customer records, which include variables like customer characteristics like gender and age, it also includes credit scores, tenure with the bank, balance, number of products, card type, estimated salary, satisfaction scores on complaint resolution, points earned, complaint status, and exit status.

To prepare the data for analysis, several transformations were performed. Categorical variables such as Gender, Location, and Card Type were transformed into numerical form. Numerical variables such as CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary, Satisfaction Score, and Points Earned were standardized to bring them to a common scale. This process helps to ensure that certain features do not dominate others due to differences in their scale.

3 DATA ANALYSIS

The data analysis was performed using Python programming language and various libraries. Pandas[7] were used for data manipulation and calculations, Matplotlib and Seaborn[1] for data visualization, and NumPy for numerical computations. These libraries allowed us to explore the dataset, answer our research questions, and gain meaningful insights. By using the capabilities of these libraries, we gained valuable knowledge about customers, complaints, credit scores, satisfaction scores, and reward systems.

3.1 Question 1

What is the proportion of the customers that are still using the banking services compared to those that have left in the period covered in the dataset? Is there a significant difference in the proportion that the bank authority should be worried about?

For our first research question, we aimed to see how many customers remained with the bank versus those who departed. Our findings showed a clear majority, 79.6%, as shown in Table 1, still choose to use the bank. This number implies a strong client foundation and satisfaction with the services provided.

Table 1: Proportions of customers

Status	Proportion
Still Using	79.6%
Exited	20.4%

However, there's another side to consider. As Table 1 shows, 20.4% of clients decided to end their banking relationship within the data's timeframe. Although this isn't a majority, it's a significant part of the client base. Their decision could come from various issues, including potential dissatisfaction with the banks offerings.

The visual representation for this research question is shown below in Figure 1, illustrating the proportions of customers still using the bank's services versus those who have exited.

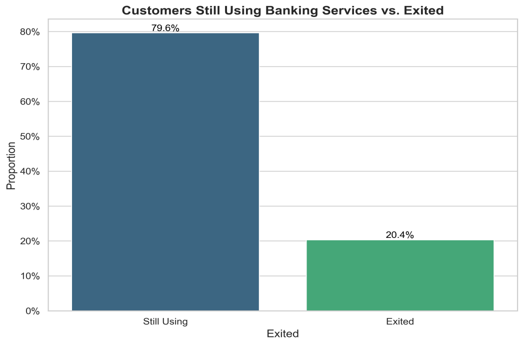

Figure 1: Still Using Banking Services vs. Exited

The bank needs to pay attention to this 20.4% of departing clients. It's crucial to identify why they've left. Was it due to service dissatisfaction, better options elsewhere, or a mix of reasons. Identifying these factors is key to reducing customer churn.

3.2 Question 2

What is the relationship between the number of complaints received by the bank authorities and the number of exited customers?

The second research question was to look into the relationship between the number of complaints received by the bank and the number of exited customers. This connection has been explored by making a crosstab that tabulates the frequency of customer exits based on whether they filed a complaint. The correlation coefficient was also calculated.

Table 2: Crosstab between Complaints and Exited Customers

Complain	**Exited=0**	**Exited=1**
0	7939	4
1	10	2027

The crosstab as we can see in Table 2 reveals that out of the customers who did not file any complaints, the majority remained with the bank while only a small number exited. However, the customers who did register complaints, the majority exited the bank, while a minimal number chose to stay with the bank. The correlation coefficient between the 'Complain' and 'Exited' variables were found to be 0.9957, indicating a very strong positive correlation.
To illustrate this relationship, a heatmap was created as seen in Figure 2, visualizing the crosstab data.

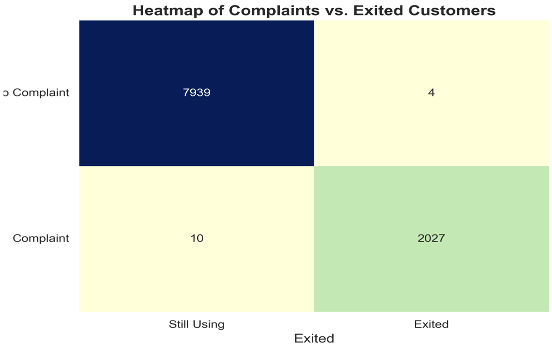

Figure 2: Heatmap of Complaints vs. Exited Customers

As seen in Figure 2 and Table 2, the analysis shows that customer complaints are a critical factor to customer churn. The bank needs to address customer complaints effectively to improve customer satisfaction and reduce the likelihood of customers exiting. This could be done by improving customer service, addressing common issues raised in complaints, and implementing measures to prevent such issues from recurring. By doing so, the bank can potentially stop the trend of customer exit associated with complaints.

3.3 Question 3

What are the characteristics and statistics (in terms of gender, age groups, and tenure) of the customers that are more likely to complain? Provide an informative profile description of those type of customers

The third research question was looking into the characteristics and statistics of customers who are more likely to complain. In this analysis, we have separated them by gender, age, and tenure to explore what types of customer who is more likely to complain.

3.3.1 Distribution of Complaints by Gender. When we look at the gender distribution of the complaining customers, we can find a noticeable gender variation. Among the customers who had registered complaints, a higher proportion were females. As seen in Table 3, the females had 1,138, compared to males who had 899.

Table 3: Number of Complaints by Gender

Gender	Number of Complaints
Female	1,138
Male	899

This shows that female customers may be more likely to file a complaint compared with males. The visual representation for this research question is shown below in Figure 3, showing the distribution of complaints by gender.

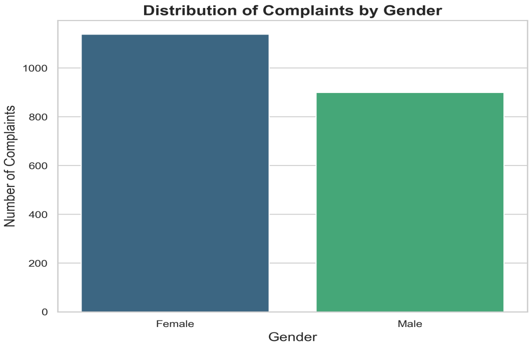

Figure 3: Distribution of Complaints by Gender

3.3.2 Distribution of Complaints by Age. For the age distribution of complaining customers, the analysis shows, as we can see in Table 4, that the majority are within the 40-49 age group, followed by the 30-39 and 50-59 age groups. The 18-29 and 60-69 age groups registered fewer complaints, and the 70+ age group filed the least number of complaints.

Table 4: Number of Complaints by Age Group

Age Group	Number of Complaints
18-29	123
30-39	479
40-49	802
50-59	484
60-69	132
70+	15

This pattern suggests that middle-aged customers, particularly those in the 40-49 age group are more likely to file complaints. The visual representation for this research question is shown below in Figure 4, showing the distribution of complaints by age.

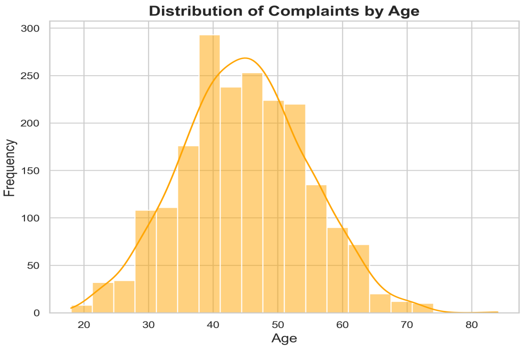

Figure 4: Distribution of Complaints by Age

3.3.3 Distribution of Complaints by Tenure. Lastly, the distribution of complaints by tenure was also analyzed for this research question.

Table 5: Number of Complaints by Tenure

Tenure	Number of Complaints
0	95
1	233
2	200
3	211
4	204
5	210
6	198
7	175
8	195
9	214
10	102

Customers who had a 1-year tenure with the bank filed the highest number of complaints, 233, closely followed by customers with a 9-year tenure, 214. The numbers then gradually decrease across the other tenure lengths, as we can see in Table 5, with customers having a 0- and 10-year tenure filing the least complaints, 95 and 102. The visual representation for this research question is shown below in Figure 5, showing the distribution of complaints by tenure.

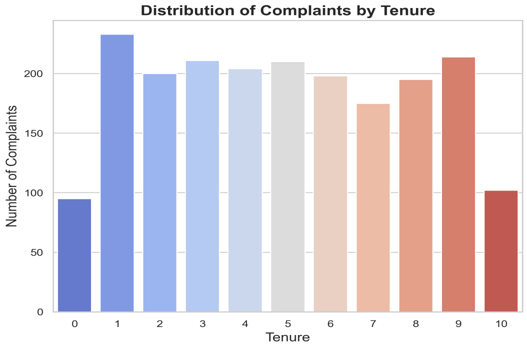

Figure 5: Distribution of Complaints by Tenure

3.3.4 Profile Description. From this analysis, when we look at these observations, a profile can be made for the type of customer who is more likely to file a complaint. They are most likely females, aged between 40 and 49, and have been with the bank for approximately one year. The bank could use this information to target that specific customer groups with improved services and complaint management strategies, to enhance customer satisfaction and reduce the number of complaints.

3.4 Question 4

Is there a significant difference between the credit scores of all the customers that have complained and those who have not in the period covered in the dataset?

For this research question, we looked at the comparison of credit scores between two distinct groups of customers, those who have registered complaints and those who haven't. Table 6 reveals a minor, yet noteworthy, difference. Customers who didn't make any complaints have a bit higher average credit score of 651.81. This is compared to the average credit score of 645.66 for customers who did make complaints.

Table 6: Credit Score by Complaint Status

Complain	Mean Credit Score	Median Credit Score
No	651.81	653
Yes	645.66	646

This trend remains consistent when we evaluate the median credit scores. Non-complainants register a median score of 653, while complainants with a median score of 646. These findings are visually represented in Figure 6, which showcases the credit score distribution among both groups via a boxplot.

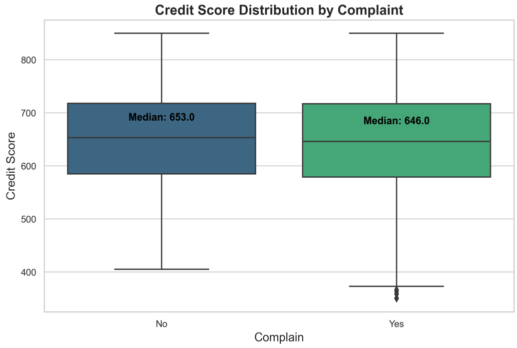

Figure 6: Credit Score Complaints

Even with these differences, the overall spread of credit scores between the two groups looks relatively similar. This leads us to consider whether the minor variations in scores between the groups truly have a substantial statistical impact. To get a clearer picture of these differences, it could be useful to dive deeper into the numbers. We could apply more robust statistical techniques, like a t-test, for instance.

While we do see a small difference in credit scores between customers who've filed a complaint and those who haven't, the difference is not substantial. Based on our analysis, it seems that credit score might not be the most reliable tool to guess whether a customer is likely to file a complaint.

3.5 Question 5

Do the satisfaction scores on complain resolution provide indication of the customers likelihood of exiting the bank?

In this research question, we analyzed whether satisfaction scores related to complaint resolution provided any indication of a customer's likelihood of exiting the bank.

Table 7: Satisfaction Score by Exit Status

Exited	Mean Satisfaction Score	Median Satisfaction Score
No	3.018115	3.0
Yes	2.995569	3.0

As we can see in Table 7, both the mean and median satisfaction scores were calculated for each group to better understand the data. The mean satisfaction score for customers who have not exited the bank is slightly higher at 3.018 than for those who have exited, at 2.996. The median satisfaction score for both groups is the same, at 3.0.

The difference in the mean scores is minimal, and the distributions are quite similar, with most scores clustering around the median value of 3. This suggests that satisfaction scores on complaint resolution might not provide a strong indication of a customer's likelihood to exit the bank.

A violin plot, as seen in Figure 7 was used to visually represent the distribution of satisfaction scores for both customers who have exited and those who are still with the bank.

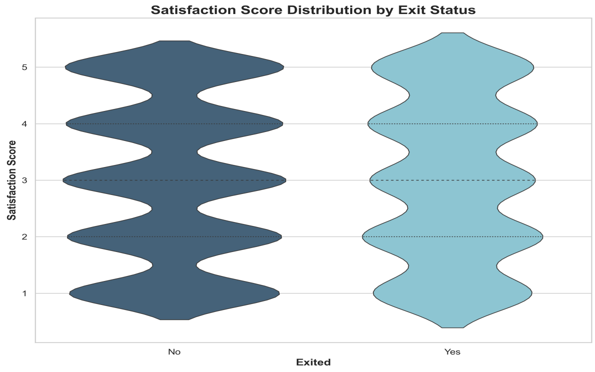

Figure 7: Satisfaction Score

While there might be a small difference in satisfaction scores between customers who exited and those who did not, the findings from this analysis do not suggest that these scores are a good predictor of a customer's likelihood to exit the bank. Other factors could potentially have a stronger influence on that decision.

3.6 Question 6

The bank has a reward system where the customers earn points when they use their Diamond, Gold, Silver, and Platinum bank card. Determine if there is a significant difference in the average points earned by the different groups of customers.

For the last research question, we are going to look into if there was a significant difference in the average points earned by customers using different card types, the card types are Diamond, Gold, Silver, and Platinum.

We calculated the mean and median points, as we can see in Table 8, earned for each card type.

Table 8: Points Earned by Card Type

Card Type	Mean Points Earned	Median Points Earned
DIAMOND	606.158210	603.0
GOLD	606.924309	603.0
PLATINUM	608.947833	607.0
SILVER	604.078778	604.5

As we can see in Table 8, the mean and median points earned don't seem to have a significant difference in the average points earned by customers using different card types. The differences are small, and all the card types have a very similar number of points, on average.

We also created a box plot, as we can see in Figure 8, to visualize the distribution of points earned by customers for each card type.

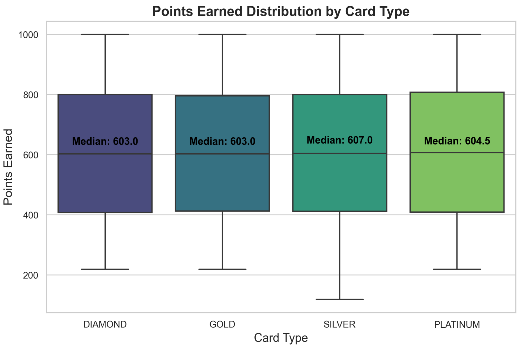

Figure 8: Points Earned Distribution

In conclusion, the bank's reward system seems to be balanced across the different card types in terms of the points earned by the customers. There may be other factors such as card benefits, limits, and fees that might play a bigger role in differentiating the card types.

4 PREDICT CUSTOMER COMPLAINTS

4.1 Model Selection and Training

The main goal of this analysis is to predict whether a customer will file a complaint or not. To achieve this, we have selected the Logistic Regression model[4]. The decision to use Logistic Regression was based on its efficiency and effectiveness in dealing with binary classification problems, which makes it well-suited to our task of predicting two possible outcomes, complaint, or no complaint[5].

Before we could start training the model, the dataset had several preprocessing steps done to it. Categorical variables, such as Gender, Location, and Card Type were transformed into numerical form. This transformation allows the model to process these variables more effectively. Then, numerical variables, such as CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary, Satisfaction Score, and Point Earned were standardized[6].

After the preprocessing part was done, the dataset was divided into two subsets, a training set, and a testing set. As we can see in Table 9, the training set, which contained 80% of the data, was used to train the Logistic Regression model. The remaining 20% of the data, the testing set, was used for evaluating the model's performance.

Data Type	Number of Records	Percentage
Training	7984	80.00%
Testing	1996	20.00%

Table 9: Training Data Split

4.2 Model Evaluation and Performance Metrics

The performance of the logistic regression model was evaluated using various metrics. Such as accuracy, precision, recall, and F1-score. Our model showed impressive results across all these metrics[8]. The accuracy score, the total predictions that are correct, was found to be 99.95%.
This high score suggests that the model is highly effective in predicting whether a customer will complain or not. The classification report can be viewed in Table 10.

	Precision	Recall	F1-score	Support
No Complaint	1	1	1	1635
Complaint	1	1	1	361
Accuracy			99.95%	1996
Macro Average	1	1	1	1996
Weighted Average	1	1	1	1996

Table 10: Classification Report

Next, we looked into the confusion matrix, as we can see in Table 11. The confusion matrix provides detailed insight into how the model is performing by revealing the types and numbers of correct and incorrect predictions. From the confusion matrix, it was observed that our model made few errors. For example, there was only one instance of a false negative, where a customer was predicted to not complain when they did.

	Predicted No Complaint	Predicted Complaint
Actual No Complaint	1634	1
Actual Complaint	0	361

Table 11: Confusion Matrix

We also analyzed the classification report we saw in Table 10. This report includes metrics like precision, recall, and F1-score. These metrics provide a more balanced view of the model's performance. It was found that the model achieved perfect scores in these metrics for both complain and not complain, further confirming its high accuracy.

4.3 Implications of Model Performance for Banking Business

Given the Logistic Regression model's high accuracy, the model can serve as a valuable tool for the bank to address potential issues with complaints. By predicting which customers are more likely to complain, the bank can address the problems that occur before it turns into a complaint and reduce customer churn[9]. This not only improves customer satisfaction but could also lead to a decrease in customers exiting the bank.

However, it is important to remember that even a highly accurate model is not without flaws and should not be relied upon without considering other factors. False positives, where the model predicts a complaint where none occurs, could lead to unnecessary interventions, and false negatives, where the model fails to predict a complaint that does occur, could result in missed opportunities to improve customer satisfaction.

Therefore, it's important that the model's performance is monitored, and the model is updated often to ensure it remains accurate in meeting changing customer behaviors and when the bank practices change.

5 PREDICTING COMPLAINTS FOR NEW SAMPLE CUSTOMERS

5.1 Data Preprocessing for New Sample

Now before we could predict complaints for New Sample customers, the New Sample dataset got a preprocessing phase similar to the Main Sample dataset. This includes encoding of the categorical variables like Gender, Location, and Card Type and standardization of the numerical variables, CreditScore, Age, Tenure, Balance, NumOfProducts, EstimatedSalary, Satisfaction Score, and Point Earned. By doing so we ensure that the new data is preprocessed in the same way that the model was trained.

5.2 Predicting Complaints using the Trained Model

After the preprocessing was complete, the Logistic Regression model that was trained on the Main Sample dataset was then applied to the New Sample dataset to predict if a customer would file a complaint or not.

5.3 Tabulating Predicted Results

For tabulating the predicted results, the predictions from the model were combined with the CustomerId from the New Sample dataset to create a table showcasing the CustomerId and their predicted complaint status, as we can see in Table 12.

CustomerId	PredictedComplaint
15710408	0
15598695	0
15649354	0
15737556	1
15671610	0
15625092	1
15741032	0
15750014	0
15784761	1
15768359	0
15805769	0
15719508	1
15609011	1
15703106	0
15626795	1
15773731	0
15756196	0
15687903	0
15777599	0
15754577	1

Table 12: Predicted Complaints for New Sample Customers

The predictions we can see in Table 12 are presented in the same order as they are in the original New Sample dataset.

6 DISCUSSION

This analysis has looked into the relationship between several variables in a bank's customer dataset to gain insights into customer complaints. The analysis also aimed to develop a predictive model for customer complaints and use it to predict customer complaints. The findings from this analysis provide noteworthy insights into the factors that contribute to customer complaints.

Firstly, the exploratory data analysis highlighted interesting trends. It was observed that females tend to register complaints more than males. It was also noted that customers between the ages of 40 and 49 filed the greatest number of complaints, with those associated with the bank for around a year also contributing significantly to the complaint numbers. This indicates that these groups might benefit from targeted customer service improvement measures.
An examination of credit scores didn't show a noteworthy difference between customers who filed complaints and those who did not. This suggests that credit scores might not serve as a reliable predictor for customer complaints. A similar trend was observed with satisfaction scores, which didn't significantly indicate a customer's likelihood to exit the bank. Even though a slight difference in scores was noted between customers who exited and those who remained, the satisfaction scores didn't conclusively determine a customer's likelihood to exit the bank.
Then, we looked into the bank's reward system. The system seemed well-balanced across the different card types that the bank offered, with no significant difference in the average points earned by customers using different card types.

The logistic regression model that was developed to predict customer complaints performed remarkably well, with an accuracy score of 99.95%. It showed excellent precision, recall, and F1-score. In spite of the high performance of the logistic regression model developed, it's important to remember potential false positives and negatives. Regular updates to the model are crucial to ensure its continued accuracy.

Despite the successful findings of this analysis, there are some areas that could be improved. For example, the analysis didn't delve into the reasons for complaints, which could provide more insights into why the complaints have been made.
Additionally, the model we made could be trained and tested with other machine learning algorithms such as decision trees[10], random forest[2], or gradient boosting[3] for comparison and potentially improve its performance. In future analyses of this bank, they could also use more factors such as transaction history, reasons for complaints, customer interactions with the bank, and a larger, more diverse sample to improve the model.

This analysis has provided valuable insights into customer complaints and has developed a model for predicting complaints. With some improvements in the future, this analysis could provide an even more comprehensive understanding of customer complaints and allow the bank to proactively address them to improve customer satisfaction and stop even more from exiting the bank.

7 CONCLUSION

This analysis has shown us various factors contributing to customer complaints within the banking sector. Through our data analysis, we found that gender, age, and tenure with the bank are significant factors in determining customer complaints. Credit scores and satisfaction scores, however, were found to be less important in determining if a customer would file a complaint or if customers would exit the bank.

The analysis also showed the effectiveness of a logistic regression model in predicting customer complaints. The model performed very well with an accuracy score of 99.95%, demonstrating its potential as a valuable tool for helping the bank reduce customer complaints and improve its complaint management.
After we applied the model to the New Sample file, we successfully predicted the likelihood of complaints. This application shows the usefulness of predictive modeling not only in understanding past patterns but also in forecasting future customer complaints.

Despite the success of this analysis, there is still room for further research and improvements. Future studies might use a broader range of factors, consider the reasons and severity of complaints, and make use of a variety of machine learning algorithms for more exact insights and potentially even more improved predictive performance.

The insights we have gained from this analysis and the predictive model that is developed could be useful in enhancing customer service and satisfaction. By understanding the factors contributing to complaints and predicting them, banks can address potential issues, improve their service, and thereby increase customer satisfaction. The results of this analysis demonstrate the value of data-driven approaches in today's competitive banking industry.

REFERENCES

[1] [n. d.]. Seaborn: statistical data visualization. https://seaborn.pydata.org/.
[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[3] Jerome H. Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics* (2001).
[4] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. 2013. *Applied Logistic Regression.* Vol. 398. John Wiley & Sons.
[5] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning.* Springer.
[6] Max Kuhn and Kjell Johnson. 2013. *Applied Predictive Modeling.* Springer.
[7] Wes McKinney. 2017. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2 ed.). O'Reilly Media.
[8] David M. Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies* (2011).
[9] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking.* O'Reilly Media, Inc.
[10] J. Ross Quinlan. 1986. *Induction of Decision Trees.*