A Personalized Overdraft Protection Framework

Karesia Ramlal and Patrick Hosein
The University of the West Indies, St. Augustine, Trinidad and Tobago
Email: karesia.ramlal@gmail.com, patrick.hosein@sta.uwi.edu

***Abstract*—Data and Artificial Intelligence are changing the business models of many financial institutions. The availability and granularity of customer data allows for the development of a personalized banking experience which has been shown to improve customer relationships and increase retention. We present a Machine Learning approach to providing personalized overdraft protection. The approach simultaneously provides benefits to both customer and bank and hence increases customer retention while improving the bank's revenue. We illustrate the approach with examples.**

***Index Terms*—Machine Learning, Overdraft Protection, Financial Decision Making**

I. Introduction

An overdraft occurs when a customer's bank account goes into a negative balance, usually as a result of processing a transaction that exceeds their available balance. Overdraft protection is offered by some commercial banks as a means of avoiding this negative balance through a pre-approved transfer of funds up to a specified limit. This service typically comes with exorbitant fees and interest rates that a customer may not be fully aware of when opting in. Those who do not opt in can be charged fees for declined transactions and non-sufficient funds (NSF). For customers already in financial difficulty, these fees are an added stress.

Existing overdraft protection options tend to be more beneficial for financial institutions than for their customers. A report published by US-based nonprofit organization, the Center for Responsible Lending (CRL), stated that commercial banks in the United States collected an estimated total of $11.68 billion USD in overdraft fees for 2019, with an average fee of $35 per overdraft transaction [1]. The organization is currently advocating for revised policy to regulate banks' overdraft practices, since vulnerable households with low balance accounts were responsible for 84% of the billions reported in annual overdraft fees [1].

A report commissioned by the Financial Conduct Authority (FCA) in the United Kingdom examined customers' attitudes towards overdraft protection and the banks that offer them [2]. The in-depth consumer research presented in the report concluded that many of the participants did not feel financially supported by banks' existing overdraft protection services. The main concern was the lack of transparency surrounding the fees and interest rates associated with overdrafts. Some participants also reported that banks were offering overdraft limits that seemed arbitrary and did not suit their individual needs.

Based on the application of machine learning techniques on customers' historical transaction data, we propose a personalized overdraft protection framework that takes these issues into account. Our data-driven solution would benefit both the customers and the financial institutions. For customers, the flexible overdraft limit would provide the desired coverage to suit their needs, and the repayment option can provide them with more control over their finances. Additionally, our approach to overdraft protection can provide more support and transparency to customers, since the features of our framework are derived from trends in the customer's banking history.

For banks, this personalized overdraft protection framework can encourage more customers to use the service as well as retain existing users. Previous research by Liu, Montgomery and Srinivasan discovered that customers were likely to close their account if they felt that the overdraft fees were disproportionate to the amount [3]. In this case, banks would lose additional revenue from all other services used by those customers. The fee structure and variable interest rates would allow the bank to provide better services to their clients, which can contribute to improved customer retention. The risk assessment component ensures that only qualified customers would be offered overdraft protection, which can eliminate the possibility of higher risk customers defaulting on their overdraft repayment.

Previous work by Wang, Cho and Denton highlighted the social value of personalization, with focus on electronic banking [4]. Their research concluded that a personalized approach had a positive effect on customer perception of online banking, which in turn led to more efficient and effective usage of online banking services. Our tailored approach is more likely to encourage customers to use the service since it is geared towards assisting customers with making better financial choices.

Our framework builds on the application of machine learning in credit risk analysis, which aims to quantify an individual's ability to repay a loan by predicting the probability of defaulting on payments. We extend this approach to overdraft protection, which can be thought of as a small short-term loan. In our case, overdraft risk refers to the customer's ability to repay their overdraft amount.

Tree-based algorithms have been shown to be more stable, have a higher accuracy, and overall better performance than deep learning models applied to the binary classification problem of predicting customer credit risk [5]. We will test two tree-based algorithms, Random Forests and XGBoost, on customer data to predict a customer's overdraft risk and whether or not they should be offered overdraft protection.

The key contribution of our model is a customer focused approach to designing overdraft fee structures, interest rates and

authorized limits. We also investigate the feasibility of deferred payments. Typically, any overdraft amount is automatically repaid when the account is credited. However, for customers in financial difficulty, the option of a deferred payment may be helpful since it would give them more control over their finances. It would also allow the bank to charge interest on the overdraft amount, which would generate revenue.

II. Related Work

A. Designing Overdraft Fee Structures

One of the key components of our personalized overdraft protection framework is the variable fee structure and interest rate. Previous work by Liu, Montgomery and Srinivasan sought to analyze and optimize overdraft fees using big data [3]. Customers were segmented into three categories based on their overdraft frequency (non-overdrafters, light, and heavy). For each segment, the researchers compared their alternative pricing strategies with the current structure of a flat per-transaction overdraft usage fee. The three alternative pricing strategies were as follows: the optimal flat fee, which was less than the original flat fee; a percentage fee based on the ratio of overdraft fee to transaction amount; and a quantity premium structure. The quantity premium fee structure was a combination of the percentage fee and the flat fee, where a customer was charged a percentage fee for the first ten overdrafts, and then a flat fee for every overdraft transaction thereafter. This strategy resulted in the highest total increase in bank revenue (5.59%).

In addition to a transaction fee, some banks charge interest on the overdraft amount used. Our framework takes into consideration personalized interest rates, which was not included in the previous research. Therefore, we present a new contribution to the design of overdraft fee structures. Our research will investigate how customer risk can be predicted using machine learning techniques, and how this predicted risk level can be used to design individualized interest rates and fees for personalized overdraft protection.

B. Machine Learning Algorithms for Credit Risk Analysis

The improved accuracy and performance of machine learning techniques over traditional statistical methods have led to many applications of artificial intelligence in the financial sector [6]. One such application is credit evaluation or credit risk assessment, which is typically modeled as a classification problem. Consumer credit risk management involves an evaluation of their ability to repay a loan, based on factors such as credit history.

In a study by Addo, Guegan and Hassani, the researchers compared the performance of machine learning and deep learning models tasked with predicting the probability of loan defaults [5]. They used logistic regression with regularization as their benchmark in order to compare the six models (random forest modeling, a gradient boosting machine and four neural networks with different criteria). The results showed though all six methods tested were an improvement on the benchmark and the gradient boosting model outperformed the other methods.

Though machine learning models are used to predict customer credit risk, the final decision is usually made by a human. Due to the nature of credit risk management, it may be necessary for financial institutions to understand the main factors that drive the model's result. This can be facilitated by explainable machine learning [7]. The researchers sought to balance the predictive accuracy of machine learning models with the degree of explanation of the results by proposing a post-processing method to interpret the output of the model. They applied the extreme gradient boosting algorithm (XGBoost) to predict the credit risk of small and medium enterprises in Europe, with two outcomes 'default' and 'non-default'. They combined XGBoost with TreeSHAP to calculate the Shapley value explanations of the companies. The results were personalized for each company and showed that the proposed explainable model identified which variables contributed most to the predicted outcome 'default'. This explainable model can contribute to a better understanding of the key factors that influence credit risk.

These studies focus on the performance and accuracy of machine learning and deep learning techniques applied to credit risk modelling. Our research builds on this by extending to the practical application of the results of these methods in designing a personalized overdraft protection framework.

C. Forecasting Customer Expenditure

One of the main concerns highlighted by customers in the FCA's market research study was the issue of arbitrary overdraft limits [2]. Our framework aims to personalize these limits to suit the customer's financial needs by predicting their expenditure using time series analysis and forecasting.

Researchers Rafi et al. used statistical models, Auto Regression Integrated Moving Average (ARIMA) and multivariate time series model Vector Auto Regressive Moving Average with Exogenous variables (VARMAX) to predict ATM cash demand for one financial institution [8]. They found that their proposed approach was an improvement upon previous implementations as their models had lower RMSE values. Our transaction dataset consists of time series data. The application of statistical models to analyse and forecast customer spending is relevant to our personalized overdraft protection framework as it can be used to predict customer's need for overdraft protection, as well as the optimal authorized limit.

D. Personalization in the Banking Industry

The scope of customer data collected by financial institutions extends beyond demographic features like age and gender. It includes behavioural data from customers' transaction history, as well as a complete view of their income, investments, loans and other banking services. Data analytics backed by artificial intelligence and machine learning can allow banks to use customer data to provide them with user specific banking experiences. Deloitte's 2021 Banking and Capital Markets Outlook cited personalized services as one of

the key factors in promoting customer satisfaction and digital engagement [9]. In our research, we explore how customer data and machine learning can be used to develop personalized overdraft protection services.

A study by Wang et al. examined the impact of personalization on electronic banking through a survey conducted in 30 branches of one financial institution in southern China [4]. The researchers concluded that personalization resulted in customers having a more favourable perception of the practicality and ease of use of the electronic banking platform. Their research demonstrated that personalization can have both practical and social implications in the banking sector, through providing insights for customer segmentation, as well as providing support for customers that are new to electronic banking.

Another study by Sunnika, Bragge and Kallio. investigated the effectiveness of personalized marketing in online banking by examining the response and behaviour of customer groups when presented with personalized messages as opposed to the default [10]. They quantified the effectiveness of the marketing strategies by measuring the pull percentage, which they defined as the number of purchases divided by the number of customers exposed to each type of campaign (personalized vs. default). The researchers found that the personalized promotion had a higher pull percentage than the existing marketing strategy.

The current research shows that customer behaviour in the banking sector can be positively influenced by personalization. Our research builds upon this by extending that personalization to financial services. Our framework focuses on overdraft protection facilities, but there is room to apply our solution to other banking services such as loans or mortgages.

III. Methodology

A. Dataset Description

We obtained two anonymized datasets from an unnamed financial institution. Our data consisted of customer profile information – 15 features describing 11,520 account holders, and transaction history – and approximately 1 million transactions from 2018-2020.

The customer profile dataset consisted of a randomly generated unique identifier (account number), demographic data (age, gender, marital status, occupation) and credit data. Similar to the credit approval approach, this dataset also included the binary target variable, where a class value of 1 indicated that overdraft protection was offered to that customer (i.e. their application for overdraft protection was approved), and a class value of 0 indicated that the customer did not qualify for overdraft protection. The transaction history dataset consisted of outgoing customer transactions over a two year period. The date, amount, and purpose for each transaction were also given. All transaction amounts were converted to USD to allow for comparison. There were 37 missing numerical values in the customer profile dataset that were replaced with the mean. The transaction history dataset was missing 41% of the values in the purpose column, since this variable did not affect the transaction amounts, it was omitted from the analysis.

After data pre-processing, we constructed additional features that described the transactional behaviour of customers. These features included average transaction amount, mean monthly expenditure, and transaction rates, derived from the customer's transaction history. We extracted the number of months that a customer was active from the date of their earliest transaction and their most recent transaction. This was used to determine monthly transaction rates and expenditure. We also derived the credit card balance/limit ratio, also referred to as utilization, which was calculated from the average unpaid balance on a client's credit card, expressed as a fraction of their authorized credit limit. Our feature engineering resulted in 5 new fields for the customer profile dataset.

B. Tree-based Algorithms for Overdraft Risk Analysis

For the purpose of our study, we assume that overdraft protection is offered on an application basis. This allows the financial institution to better manage their risk, as well as for the customer to be familiar with the terms and conditions of overdraft protection. We express overdraft approval and overdraft default risk as two binary classification problems. For predicting overdraft approval, the target variable is divided into two classes – overdraft granted (class value 1) and overdraft not granted (class value 0). For predicting overdraft default risk, the target variable also has two class labels – default (class value 1) and non-default (class value 0). We model the customer's overdraft default risk based on their credit card payment habits. We assume that their ability to pay off their credit card debt is indicative of their ability to repay an overdraft amount.

The classifier predicts the probability of the customer's overdraft application being approved (overdraft granted). This is similar to the approach of using binary classification for credit risk analysis. The model also gives the predicted likelihood of a customer repaying their overdraft amount. For example, a customer's overdraft application was predicted as accepted (class value 1), and non-default (class value 0) with a probability score of 0.15, which means that the customer has an 85% chance of repaying their overdraft amount, and 15% chance of defaulting.

We define overdraft risk as the probability of a customer defaulting on their overdraft repayment. For the case of overdraft protection, the overdraft amount is automatically deducted from your account when the funds are available. A customer that defaults on their overdraft payment is one who uses up their overdraft limit, but does not make any deposits to the account after, thus resulting in a loss for the financial institution. We consider two algorithms, random forest and XGBoost for the task of modeling customer overdraft approval and default risk.

1) Random Forests: By definition, random forests are an ensemble of decision trees where the growth of each tree is dependent on a random vector that is independent and identically distributed for all tree-structured classifiers in the forest

[11]. The bootstrap aggregation ensures that the individual trees in the random forest have low correlations with each other. Similar to bagging, the decision trees in the random forests are grown on bootstrapped samples in order to reduce variance. In order to classify a new instance, each decision tree classifier votes for a class label and the majority class is chosen as the predicted label. Random forests are robust to noisy data and outliers. They also do not overfit, since the law of large numbers ensures that the generalization error converges even when more trees are added [11]. The performance of random forests is better than that of individual decision tree classifiers, and comparable to boosting ensemble methods. Random forest classifiers are fast and especially efficient on large datasets since they consider fewer attributes for each split [12]. One disadvantage of random forest classifiers is that the models are not easy to interpret and sometimes treated as a black-box [13].

2) Extreme Gradient Boosting Algorithm (XGBoost): The second tree-based classifier evaluated for overdraft risk modelling in our proposed framework is the extreme gradient boosting algorithm or XGBoost, which is based on the boosting ensemble method. XGBoost is a scalable system of gradient boosted decision trees [14]. Gradient boosted trees are an additive model of decision trees that use gradient descent optimization to minimize a differentiable loss function. According to XGBoost creator, Tianqi Chen, its scalability is due to its sparsity aware tree learning algorithm and parallel and distributed computing. The algorithm is designed for speed and efficiency, and has consistently outperformed other methods across a wide range of problems, as demonstrated by its success in machine learning and data mining competitions [14]. XGBoost has also been combined with Shapley value explanations which resulted in an interpretable machine learning model.

C. Time Series Forecasting for Determining Overdraft Limits

After predicting a customer's overdraft application approval, and their risk of default, we performed time series modelling on their average monthly expenditure to determine what would be a feasible overdraft limit for that customer. Modelling customer expenditure is a univariate time series problem, as the transaction amount is the only time dependent variable. Due to the large number of customers, it is infeasible to model each one individually. Instead, we model the time series forecasts for customer segments. Additionally, we consider a confidence interval for the forecasted limits.

1) Simple Exponential Smoothing: We did not observe any clear trends or seasonality in our data, as such we selected the simple exponential smoothing (SES) model, which is commonly used for data without trends. SES is a time series forecasting model that uses the concept of a weighted average and assigns lower weights to older observations. The formula is given by:

$$S_t = \alpha y_{t-1} + (1-\alpha) S_{t-1} \tag{1}$$

where α is the smoothing constant, $0 < \alpha < 1$ and t is a time period. The smoothing parameter, α controls the weight assigned to recent observations. The value of α is chosen to minimize the mean squared error, which is a measure of the model's fit, given by the equation:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2}{N}} \tag{2}$$

where x_i is the actual time series observation, $\widehat{x}_i$ is the estimated time series and N is the number of data points.

D. Optimizing Overdraft Fee and Interest Rate

We propose a stratified fee structure where a flat fee is charged per overdraft transaction up until a threshold value. When the number of overdraft transactions exceeds the threshold, a higher fee is charged, corresponding to the customer's risk level. Based on the transaction rates of the customers in our dataset, we set the threshold value for overdraft frequency at 5 transactions.

We focus instead, on the determination of a variable interest rate. We make the assumption that a customer with insufficient balance to cover their transactions would incur a penalty for that declined transaction. We called this penalty the cost of missed payment. If that customer had overdraft protection, they would pay a fee for their transactions, as well as an interest rate on their authorized limit. We called this amount the cost of overdraft protection, but it is also equal to the revenue gained by the bank as a result of offering overdraft protection to that customer.

The individualized interest rate was based on the costs of overdraft protection vs. the cost of missed payment. A customer's net savings would be the cost of missed payment minus the cost of overdraft protection. The trivial solution to maximizing customer net savings would result in a 0% overdraft interest rate, which would mean zero revenue generated and is therefore infeasible. The other extreme would be to maximize bank revenue (cost of overdraft protection), which would not benefit the customer. Since our interest rate needs to benefit both the customer and the bank, we used linear interpolation to determine the optimal value.

1) Linear Interpolation: Linear interpolation is a numerical method for estimating the value of a function, assuming that it lies on a straight line between two consecutive data points [15]. It is useful for our purpose of finding the optimal interest rate since the solution lies on the line between the cost of overdraft and the cost of missed payment. Given two points on a line (x_i, y_i) and (x_{i+1}, y_{i+1}), and a point x in between them, ie. $x_i < x < x_{i+1}$, the linear interpolation at x or value of $y(x)$ is given by:

$$\widehat{y}(x) = y_i + \frac{(y_{i+1} - y_i)(x - x_i)}{(x_{i+1} - x_i)} \tag{3}$$

where x_i and x_{i+1} are the minimum and maximum cost of missed payment respectively, and y_i and y_{i+1} are the maximum and minimum cost of overdraft protection for each customer. The value of $\widehat{y}(x)$ was used to determine the optimum interest rate for individual customers.

E. Survival Analysis for Determining Deferral Period

Our personalized overdraft protection framework incorporates the option of deferred payments. Typically, overdrafts are automatically repaid once the account balance is credited. For individuals in financial difficulty this automatic payment may cause them to overdraft again, which can be a frustrating experience and lead to customer attrition. We propose the option of deferred overdraft repayment as a means of improving customer experience.

We recognize that a deferred payment may introduce an added level of risk for the financial institution, as there may be some clients that would potentially exploit this option and purposely default on their overdraft amount. We determined an optimized repayment period that comes into effect after the customer account is credited, but before the automatic payment is deducted. During this period, the bank would impose additional charges, and the customer would be able to postpone their payment up to a maximum number of days. To achieve this, we modelled the time to repayment using survival analysis.

Survival analysis is a statistical method that analyzes time to event data. While this method has its origins in medical sciences, it has been applied to various industries including forecasting machinery failure [16], determining insurance premiums [17], and predicting financial distress [18]. In our case, the event is customer repayment of their overdraft loan.

Since overdrafts don't have a stipulated due date like credit cards or loans, they may have a more open ended time to payment. However, it is not feasible to offer a deferral period of, say, 6 months. The current repayment method is via an automatic deduction from the customer's bank account as soon as funds are available, so the customer does not decide when the payment is made. To reiterate, our event is defined as the time to a *customer* repaying their overdraft amount. Since deferred payment is not an existing feature of overdraft protection, the data we need is not available. Instead, we modelled our deferred repayment option in a similar manner to the grace period between the end of credit card billing cycles and when the bill is paid. We used a non parametric method for estimating the survival function.

1) Kaplan-Meier Estimator for the Survival Function: The product-limit (PL) method developed by Kaplan and Meier is one of the non-parametric methods for estimating survival functions, meaning that it estimates the survival curve without a known probability distribution [19]. The Product-Limit estimator, or Kaplan-Meier estimator of a survival function is given by:

$$\widehat{S}_{KM}(t) = \prod_{i:\, t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{4}$$

where n_i is the number of customers present at time t_i and d_i is the number of events (number of repayments) at time t_i. The survival curve is a plot of $S(t)$ vs t and it is a step function with jumps at the times when an event is observed, and constant between the time of two events. A censored survival time occurs when the event is not observed during the specified time period of the study [20]. One of the key assumptions of the Kaplan-Meier method is that the reason for censoring is unrelated to or independent of the event, however there were no censored observations in our data.

TABLE I
PERFORMANCE METRICS FOR CLASSIFICATION MODELS – PREDICTING OVERDRAFT APPROVAL

Model	Accuracy	Precision	Recall	AUC
Random Forest	70.94	0.741	0.793	0.689
Random Forest (Tuned)	71.63	0.739	0.814	0.692
XGBoost	70.59	0.731	0.806	0.681
Logistic Regression	68.92	0.696	0.837	0.657

TABLE II
PERFORMANCE METRICS FOR CLASSIFICATION MODELS – PREDICTING OVERDRAFT DEFAULT RISK

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (Tuned)	93.10	0.980	0.906	0.941
XGBoost	91.95	0.979	0.887	0.931

IV. NUMERICAL RESULTS

A. Computing Overdraft Approval

In addition to the chosen tree-based classification algorithms – Random Forests and XGBoost, we evaluated the performance of a logistic regression model on our dataset. We also compared the performance of the random forest classifier with tuned parameters to that of the other two algorithms. The parameters for the tuned random forest classifier were determined using k-fold cross validation (where $k = 5$). The results are summarized in Table I. All 3 tree-based models had a higher accuracy than the logistic regression model, with the tuned RF model performing the best at 71.6% accuracy, and a recall of 0.8. Out of a test set of 2880 instances, with 1728 positive classes, the tuned random forest model resulted in the lowest number of false negatives. That model also had the highest AUC (0.69).

B. Computing Overdraft Default Risk

The overdraft default dataset was unbalanced, with only 2% of customers previously defaulting (class value 1). As a result, we re-sampled with SMOTE [21] after splitting into training and testing sets. We then applied the same algorithms (Random Forests and XGBoost) to our overdraft default dataset to predict each customer's risk of defaulting on a payment. The performance of these models is summarized in Table II. Both models had a high accuracy, but since the classes were imbalanced, we calculated the F1 score, which is a harmonic mean of the precision and recall metrics. Both models had high F1 scores, indicating that they performed well for the given risk classification task.

The feature importance bar chart obtained from the XGBoost classifier is shown in Figure 1. Features 2 and 8 have the most influence on the model. These features correspond to the customer age and their average transaction amount.

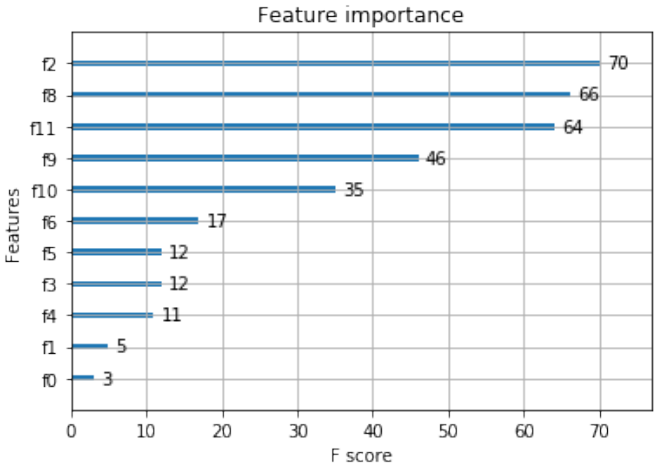

Fig. 1. XGBoost Feature Importance

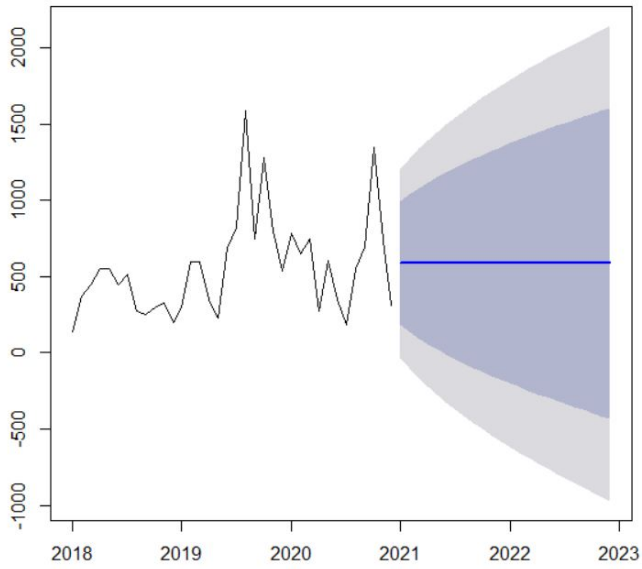

Fig. 2. Prediction Intervals for Forecasted Authorized Limits

Another important feature was average monthly expenditure (feature 11). While two of the top 3 features are transaction related, individuals with shorter transaction history can still be included in the overdraft default risk model since the most important feature was derived from demographic information.

C. Time Series Forecasting

If the models predicted that the customer's overdraft application was accepted, and that their risk of default was ≤ 0.5, their authorized limit was calculated. Our time series model determined a prediction interval, which was an upper and lower limit for the customer's forecasted expenditure. Given the number of customers, it was not feasible to do individual time series for all of them. Instead, we modelled the personalized limits based on customer segments. We separated the customers into two segments, low and high volume, based on their transactional behaviour. The RMSE for the SES models with $\alpha = 0.8$ and $\alpha = 0.9$ were 0.87 and 1.35 for the low and high transaction volume customers respectively. These values indicate that the exponential model performed relatively well when tasked with forecasting monthly expenditure.

We also separated customers into segments based on their demographic data. Figure 2 shows the prediction interval for the forecasted authorized limit for a low transaction volume customer with the following characteristics: age – 31, marital status – single, gender – male, occupation – engineering & technology. The blue line is the forecast, the purple and grey shaded areas are the 80% and 95% prediction intervals respectively. From this plot we can see than an authorized limit of $500 would be suitable for this customer. For individuals with an insufficient number of transaction periods, their overdraft limit would be based on their customer demographics segment.

D. Fee Structure and Interest Rates

Our stratified fee structure is shown in Table III. Previous studies in designing overdraft pricing strategies evaluated the option of fees that were proportional to the overdraft amount. We apply this approach of percentage fees to our data, however, we determined the fees based on customer risk and usage.

TABLE III
PROPOSED OVERDRAFT FEE STRUCTURE

Customer Risk	1-5 Transactions	6+ Transactions
Low	6.35	9.92
Medium	12.25	16.89

The fee amounts can be used as is or rounded up to the nearest dollar to prevent confusion when presenting the fee structure to customers. We segmented our customers into two groups based on their predicted default risk. These groups were further subdivided into two more categories based on their transaction rates. Individuals with a transaction rate less than 10, were classified as low volume, and the other group was labelled high volume. The percentage fees were then calculated for each of these four groups, based on their default probability and their average transaction amount. The mean percentage fee for each group was set as the overdraft usage fee for each group. This method ensures that financial institutions charge a fair fee for overdraft usage.

We calculated the individual interest rate based on linear interpolation at the median of the line between the maximum cost of missed payment denoted by C_{MP} and the maximum cost of overdraft protection denoted by C_{OD}. The cost of missed payment was determined using the penalty fees a customer would incur by missing a third party loan or credit card payment due to inadequate account balance. For customers where there was no information available on third party payments, the null values were filled with the modal value for

that customer risk segment.

$$C_{OD} = (\text{interest rate} \times \text{authorized limit}) + \text{usage fee} \quad (5)$$

$$C_{MP} = (\text{third-party int rate} \times \text{amount due}) + \text{NSF fee} \quad (6)$$

The results from the risk classification using XGBoost suggested that age had a significant effect on a customer's predicted default risk. For a sample of customers, we determined their personalized interest rate vs their age and groups. We found that customers between the ages of 30-40 with a medium risk level had higher interest rates. Most of the older customers (60+) had a medium risk level but their interest rates were generally below 16%.

E. Deferred Repayment Period

The survival curve generated from Kaplan-Meier estimates of survival time is shown in Figure 3. We can see that the median survival time is approximately 31 days, meaning that at least 50% of customers take longer than 31 days to repay. The length of the deferral period is directly proportional to the revenue generated from interest on the overdraft amount. However, as the deferral period increases, the risk of default also increases. From the survival estimates, we see that at Time = 31 days approximately half the number of all customers chose to make a payment. By 45 days, approximately 85% of the customers made a payment.

Based on customer behaviour with repaying their credit card balance, we can conclude that, if a customer was given the choice of when to repay their overdraft limit, 50% of them would repay before 31 days. Therefore, we propose a deferral period of 31 days, during which the bank would accumulate interest on the customer's overdraft amount. This deferral period would come into effect once their account balance has been recharged. If the customer fails to voluntarily repay their overdraft amount within the stipulated 31 days, then the bank would automatically deduct the outstanding amount. This deferred repayment is intended to provide customers with more flexibility and control when it comes to managing debt.

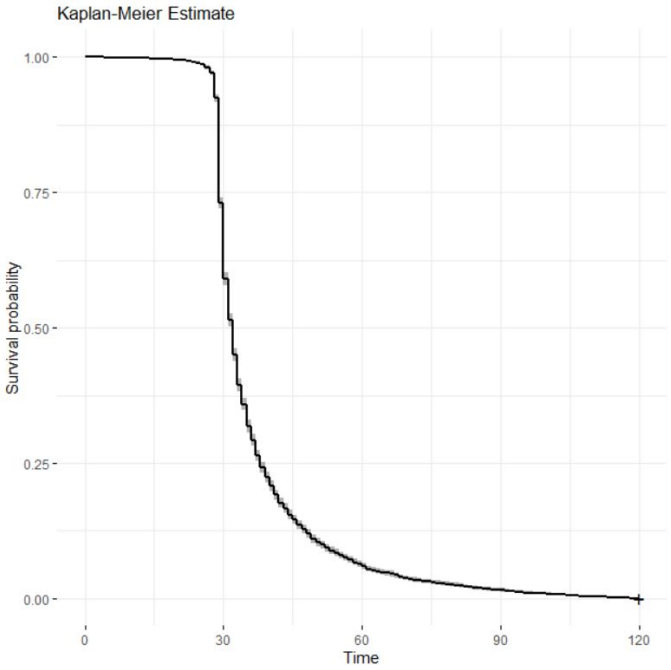

Fig. 3. Survival Function for Time to Repayment

TABLE IV
COMPARISON OF PROPOSED AND EXISTING FRAMEWORK

Customer	1	2	3	4	5
Limit	920	1900	750	515	590
Interest (old)	10	10	10	10	10
Interest (new)	8	14	13	16	12
Fee (old)	9	9	9	9	9
Fee (new)	6.35	6.35	6.35	12.25	12.25
Revenue (old)	101	199	84	60.5	68
Revenue(new)	80	272	104	95	83
Difference	-21	73	20	30.5	15

V. DISCUSSION

From our initial dataset of 11520 customers, we focused only on the subset of who were approved for overdraft protection (approximately 6000 customers), since they would be the ones who would benefit from our personalized framework. In order for our solution to be implemented, there must be some incentives for the financial institution to consider modifying their service. Other than improving customer retention and loyalty, there are some quantifiable benefits of our framework.

We compare the revenue generated from existing overdraft protection offerings to that of our proposed framework for 5 customers. We chose to compare the two options on the customer level, rather than on a cumulative level, to allow for a more granular perspective that can be scaled up to estimate the revenue generated for the entire customer base. Also, it would be misleading to assume that all of the customers who are offered overdraft protection would use it. Therefore, to avoid these assumptions, we compare the two frameworks for typical banking customers. These are outlined in Table IV.

By implementing personalized overdraft protection with interest rates and authorized limits based on customer data, the financial institution would generate a net increase in revenue of $121.50 from these 5 customers. Another way banks can benefit from our proposed framework is through the implementation of a deferred repayment period, where extra interest would accumulate on the customer's overdraft amount over a period of 31 days.

Debt management continues to be a challenge for low and middle income families. Personal debt and financial distress have even been linked to mental health problems [22]. There is a lack of investigation into how the average person can overcome financial difficulty. Our proposed personalized overdraft protection framework was designed with the average customer in mind. Though overdrafts are common in the

corporate banking sector, we chose to focus on how it can be optimized for personal banking. We expect that by providing overdraft protection with variable interest rates, a stratified fee structure, and deferred payment, our proposed overdraft protection framework would benefit customers in financial difficulty or those faced with sudden expenses. We expect that the customer benefits included in our framework would encourage existing clients to use the service more frequently, as well as attract new clients to the bank. Customer satisfaction and changes to customer retention are two metrics that can be used to monitor the performance of the framework.

Financial institutions have been collecting data on their customers for a long time. With the recent shift to online platforms, more of this data is now in a form that can be analysed using machine learning techniques to provide meaningful insights about their customer base that can lead to personalized services. Furthermore, the use of local or regional data allows for location specific recommendations, and the development of well-trained models.

The positive social effects of personalization include an improvement in customer perception of banking services. Our personalized approach to overdraft protection is intended to enhance the customer banking experience. The social impact of personalized overdraft protection can be evaluated through customer feedback. We conducted our own market research to ascertain preliminary reactions to the framework, and 83% of responders favoured personalized banking services.

VI. Conclusion

By using a combination of machine learning techniques, statistical methods and numerical methods, we were able to design a personalized overdraft protection framework that maintained or improved both customer welfare and bank revenue. The personalized features were based on market research and the literature related to consumer opinions on overdraft protection. Additionally, the comparison of our proposed solution to the existing frameworks showed that financial institutions would not incur any significant losses in revenue by implementing our personalized framework.

The use of customer data to personalize their banking experience can also extend to monitoring and forecasting customer deposits and expenditures which can be used to set up balance alerts or potentially predict overdrafts before they occur. Monitoring the fluctuations in their account balance over time can provide insight as to which banking products and services a customer would benefit from, or be likely to adopt, and this can be achieved by way of a recommender system. As more financial institutions encourage the adoption of online and digital channels, the quality and availability of customer data will vastly improve and this will contribute to the development of new models.

Our research applied the core concepts of data science to solve a problem with both social and economic impact. Our holistic approach focused on the performance of the techniques as well as the feasibility of our solution. We demonstrated the advantages of our framework for customers, in the form of net savings, and we identified the potential for banks to generate additional revenue through deferred repayment. Our work on overdraft protection can be used as a starting point for fair usage policy on regulating overdraft practices.

References

[1] S. B. Peter Smith and R. Borne, "Banks must stop gouging consumers during the covid-19 crisis. overdraft fees," https://www.responsiblelending.org/sites/default/files/nodes/files/research-publication/crl-overdraft-covid19-jun2019.pdf, 2019, [Online; accessed June-2021].

[2] Atticus, "Consumer research on overdrafts report," https://www.fca.org.uk/publication/research/consumer-research-on-overdrafts.pdf, 2018, [Online; accessed June-2021].

[3] X. Liu, A. Montgomery, and K. Srinivasan, "Analyzing bank overdraft fees with big data," *Marketing Science*, vol. 37, no. 6, pp. 855–882, 2018. [Online]. Available: https://doi.org/10.1287/mksc.2018.1106

[4] M. Wang, S. Cho, and T. Denton, "The impact of personalization and compatibility with past experience on e-banking usage," *International Journal of Bank Marketing*, vol. 35, pp. 45–55, 2017.

[5] P. M. Addo, D. Guegan, and B. Hassani, "Credit risk analysis using machine and deep learning models," *Risks*, vol. 6, no. 2, 2018. [Online]. Available: https://www.mdpi.com/2227-9091/6/2/38

[6] A. Bahrammirzaee, "A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems," *Neural Computing and Applications*, vol. 19, pp. 1165–1195, 2010.

[7] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Computational Economics*, vol. 57, no. 1, pp. 203–216, 2021.

[8] M. Rafi, M. T. Wahab, M. B. Khan, and H. Raza, "Atm cash prediction using time series approach," in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2020, pp. 1–6.

[9] M. Shilling and A. Celner, "2021 banking and capital markets outlook," https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-outlooks/banking-industry-outlook.html, 2021, [Online; accessed June-2021].

[10] A. Sunikka, J. Bragge, and H. Kallio, "The effectiveness of personalized marketing in online banking: A comparison between search and experience offerings," *Journal of Financial Services Marketing*, vol. 16, no. 3-4, pp. 183–194, 2011.

[11] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[12] J. Han, M. Kamber, and J. Pei, "Data mining concepts and techniques third edition," *The Morgan Kaufmann Series in Data Management Systems*, vol. 5, no. 4, pp. 83–124, 2011.

[13] J. VanderPlas, *Python data science handbook: Essential tools for working with data*. " O'Reilly Media, Inc.", 2016.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[15] T. Siauw and A. Bayen, *An introduction to MATLAB® programming and numerical methods for engineers*. Academic Press, 2014.

[16] A. Heng, A. C. Tan, J. Mathew, N. Montgomery, D. Banjevic, and A. K. Jardine, "Intelligent condition-based prediction of machinery reliability," *Mechanical Systems and Signal Processing*, vol. 23, no. 5, pp. 1600–1614, 2009.

[17] D. Sari, D. Lestari, and S. Devila, "Pricing life insurance premiums using cox regression model," in *AIP Conference Proceedings*, vol. 2168, no. 1. AIP Publishing LLC, 2019, p. 020034.

[18] E. K. Laitinen, "Survival analysis and financial distress prediction: Finnish evidence," *Review of Accounting and Finance*, 2005.

[19] V. Bewick, L. Cheek, and J. Ball, "Statistics review 12: survival analysis," *Critical care*, vol. 8, no. 5, pp. 1–6, 2004.

[20] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[21] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[22] C. Fitch, S. Hamilton, P. Bassett, and R. Davey, "The relationship between personal debt and mental health: a systematic review," *Mental Health Review Journal*, 2011.