ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ

CREDIT RISK ANALYSIS VIA MACHINE LEARNING METHODS: CLIENT SEGMENTATION BASED ON PROBABILITY OF DEFAULT

Kassiani Grigoriou

Supervising Professor: Dr. Dasilas Apostolos

Examining Committee: Dr. Refanidis Ioannis & Dr. Steiakakis Emmanouil

Objective

“A client segmentation based on probability of default using machine learning techniques is the aim of this dissertation.”

Introduction

- The rapid evolution of the technology, the competitive environment, as well as the huge amount of data that is available today, lead businesses to switch to the new digital reality.

- Automation of processes and data driven decision-making using new methods such as artificial intelligence and machine learning are a primary goal of organizations.

Introduction

- Credit risk is considered as the most important risk for a financial organization.
- For over two hundred years, in the field of prediction of the bankruptcy of an organization most evaluations were done subjectively. [Bellovary et. al. (2007)]
- According to *McKinsey & Co*, the risk functions in banking institutions should be very different by 2025.

Introduction

Machine learning attribution to the risk management field could be extremely high in the next few years, as it:

- ✓ has the ability to analyze large volumes of data
- ✓ can identify complex, linear and non-linear patterns
- ✓ can be continuously trained and improved
- ✓ can lead to accurate risk models

Introduction

Applied fields that Machine Learning can be used in the financial sector:

- Risk management
- Fraud detection
- Credit risk modeling
- Money laundering
- Behavior monitoring
- Customer support

Introduction

- According to *Malhotra* (2003) on the subject of credit risk analysis, ML techniques are superior to traditional statistical models.

- Credit scoring using machine learning is generally done using some kinds of classifier that differentiates between trusted and unreliable customers using previous customer data.

- Neural Networks, SVM, Naive Bayes, Bayesian Networks, Decision Tree, Random Forest, Hybrid and Ensemble models are a few ML techniques used by researchers for credit scoring.

Data & Methodology

- The dataset that is used in this thesis can be found on Kaggle.
- It contains 32,581 borrowers' data.
- There are 11 variables per borrower, which are the following:

  Age, Annual income, Home ownership, Employment length, Loan intent, Loan grade, Loan amount, Interest rate, Loan status, Percent income, Historical default

Data & Methodology

- Python programming language and scikit library have been used for the data analysis and models' development.
- Supervised learning has been applied. An input and an output variable are necessary, whereas a training set based on predefined inputs and outputs is used for teaching the models to predict the correct output.
- 12 different models have been deployed.
- Decision tree, Random Forest and KNN (k-Nearest Neighbors) Machine Learning algorithms have been developed.

Data & Methodology

- Label encoding - a machine learning encoding technique has been done to convert categorical and text data (columns) to numerical.

- The dataset is split in two parts, the train set which contains the 80% of the initial dataset and the test set which contains the rest 20% of the observations.

Data & Methodology

For each algorithm used, 4 models have been created.

1. Initial Model – Default parameters

2. Model with best parameters – They are identified with the Randomized Search CV technique of hyper tuning procedure.

3. Model after feature importance implementation

4. Model using SMOTE method – an oversampling technique to handle imbalanced data.

Data & Methodology

Models' evaluation is based on the following metrics:

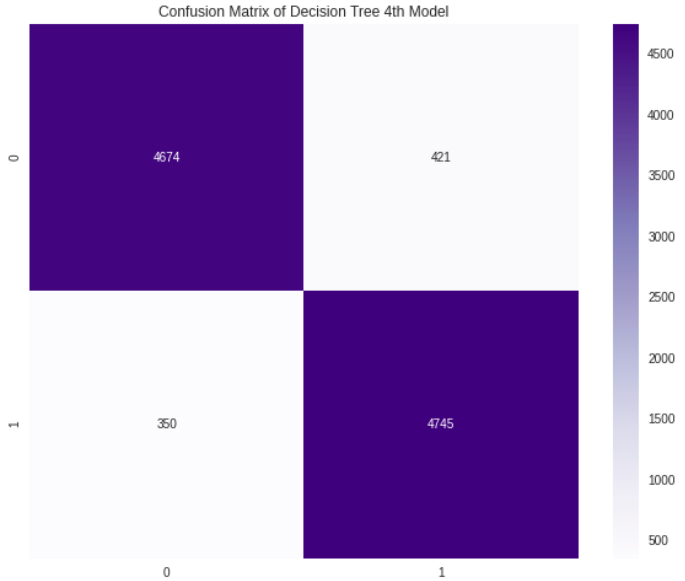
1. Confusion matrix
2. Precision and Recall
3. F1 Score
4. ROC-AUC Score and ROC Curve

Empirical Results

Decision Tree

Decision Tree – It was concluded that the best model was built on Smote dataset.

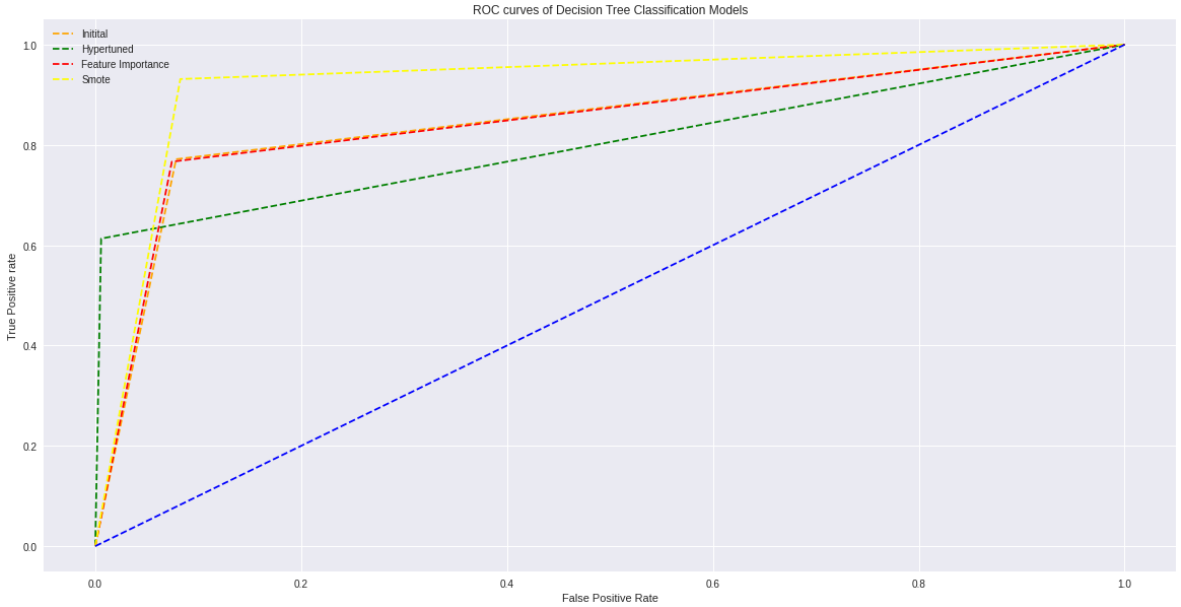
- From 10,190 people, 9,419 were predicted in the correct class.
- 4,674 people were classified in class 0, which means that they would take a loan and that was correct as they did take the loan
- 4,745 classified in class 1, which means they would not take a loan and actually they did not take it.



Empirical Results

Decision Tree

Scores	Initial model	Hypertuned model	Feature Selected model	Smote Data model
Accuracy	88.8752%	**91.1155%**	89.0900%	**92.4337%**
ROC-accuracy	84.6470%	80.3764%	84.6070%	924337%

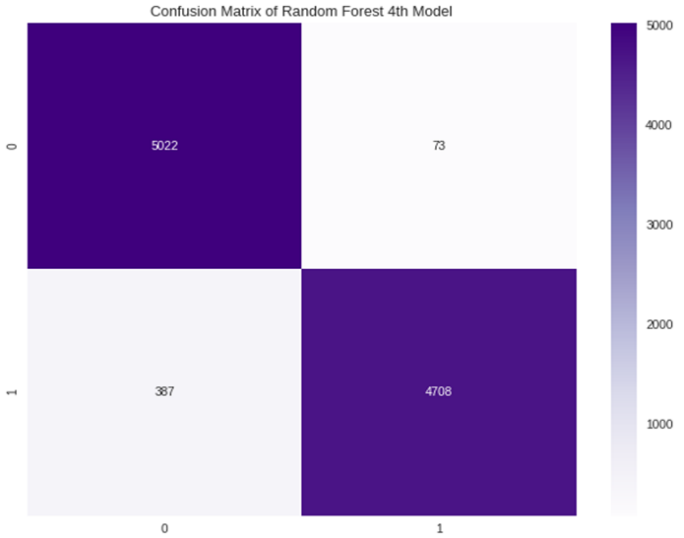


Empirical Results

Random Forest

Random Forest – It was concluded that the best model was built on Smote dataset.

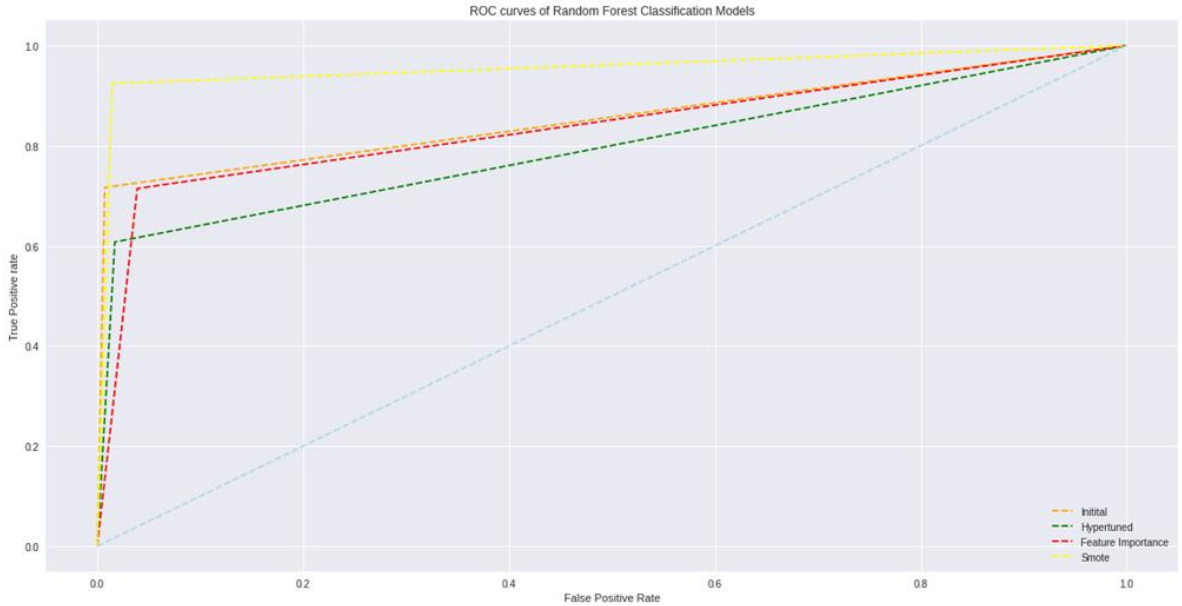
- From 10,190 people, 9,730 were predicted in the correct class.
- 5,022 people were classified in class 0, which means that they would take a loan and that was correct as they did take the loan
- 4,708 classified in class 1, which means they would not take a loan and actually they did not take it.



Empirical Results

Random Forest

Scores	Initial model	Hypertuned model	Feature Selected model	Smote Data model
Accuracy	93.2484%	**90.1028%**	90.7320%	**95.4858%**
ROC-accuracy	85.5467%	79.5260%	83.7812%	95.4858%



Empirical Results

KNN

KNN – It was concluded that the best model was built after Hypertuning implementation.

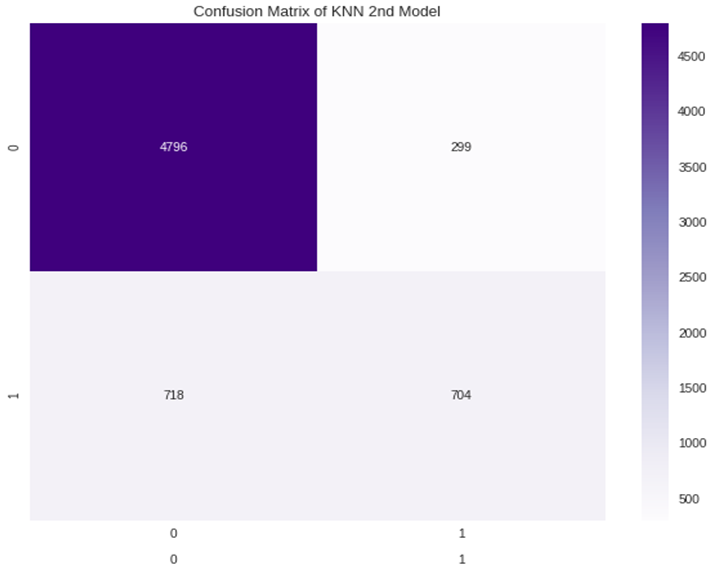
The best parameters that were resulted from the Hypertuning technique and *Grid Search* were:

- Distance metric: '**Euclidean**'
- Power parameter for the Minkowski metric **p**: **2**
- Best number of neighbors (**n_neighbors**): **13**

Empirical Results

KNN

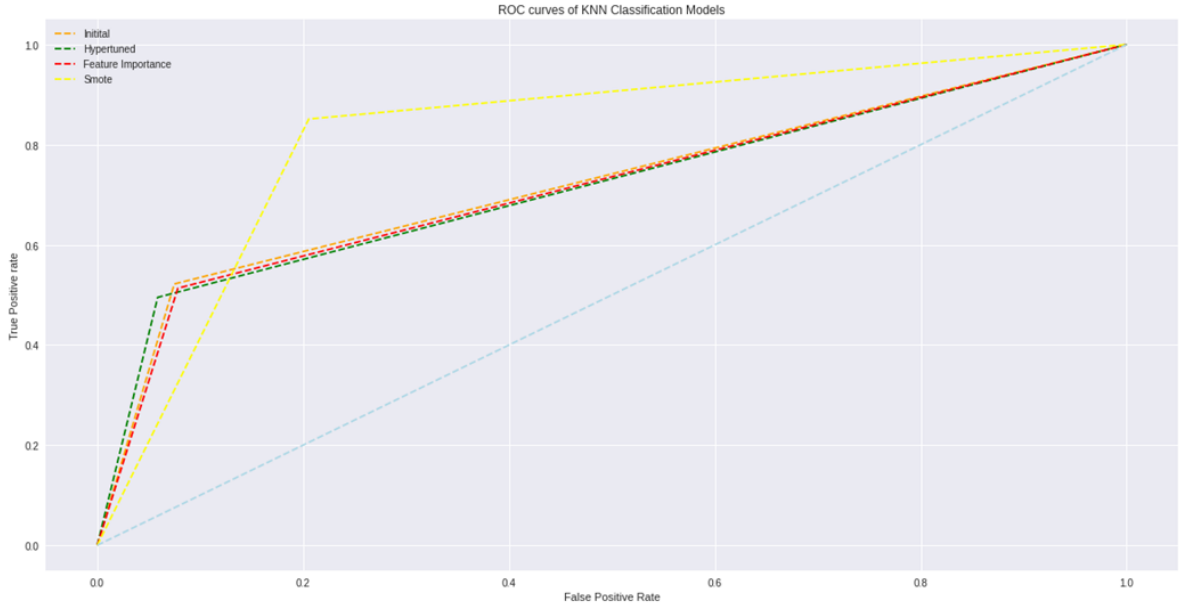
- From 6,517 people, 5,500 were predicted in the correct class.
- 4,796 people were classified in class 0, which means that they would take a loan and that was correct as they did take the loan
- 704 classified in class 1, which means they would not take a loan and actually they did not take it.



Empirical Results

KNN

Scores	Initial model	Hypertuned model	Feature Selected model	Smote Data model
Accuracy	83.7195%	**84.3947%**	83.2592%	**82.2865%**
ROC-accuracy	72.3510%	71.8196%	71.7525%	82.2865%

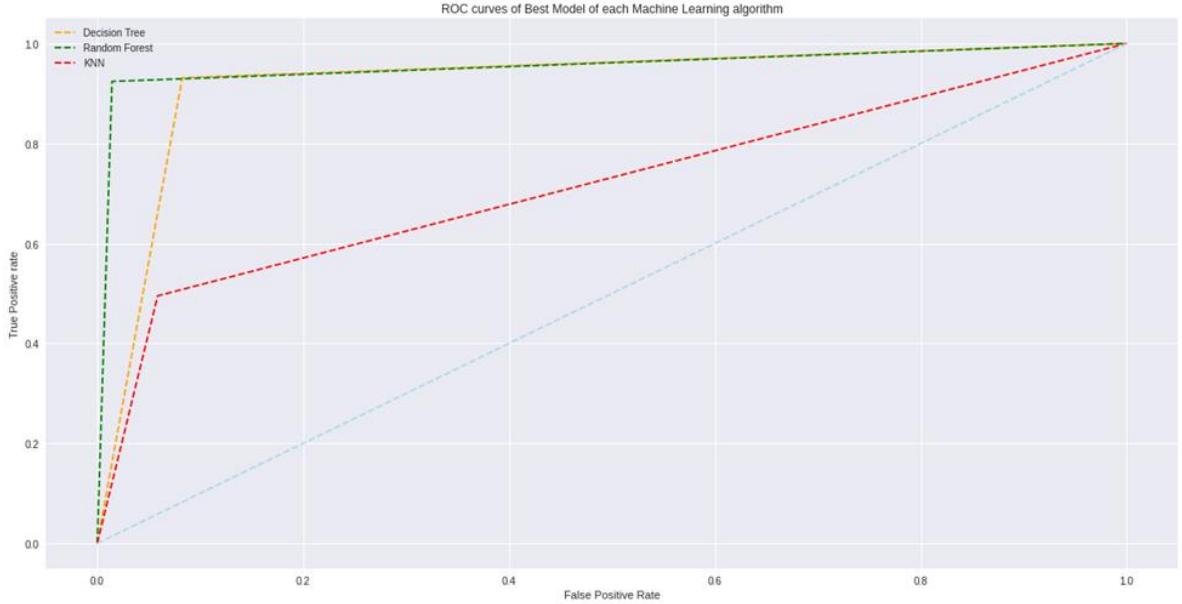


Empirical Results

Comparison

Comparison Results for the best model of each machine learning algorithm

Scores	KNN	Decision Tree	Random Forest
Accuracy	84.3947%	**92.4337%**	**95.4858%**
ROC-accuracy	71.8196%	92.4337%	95.4858%

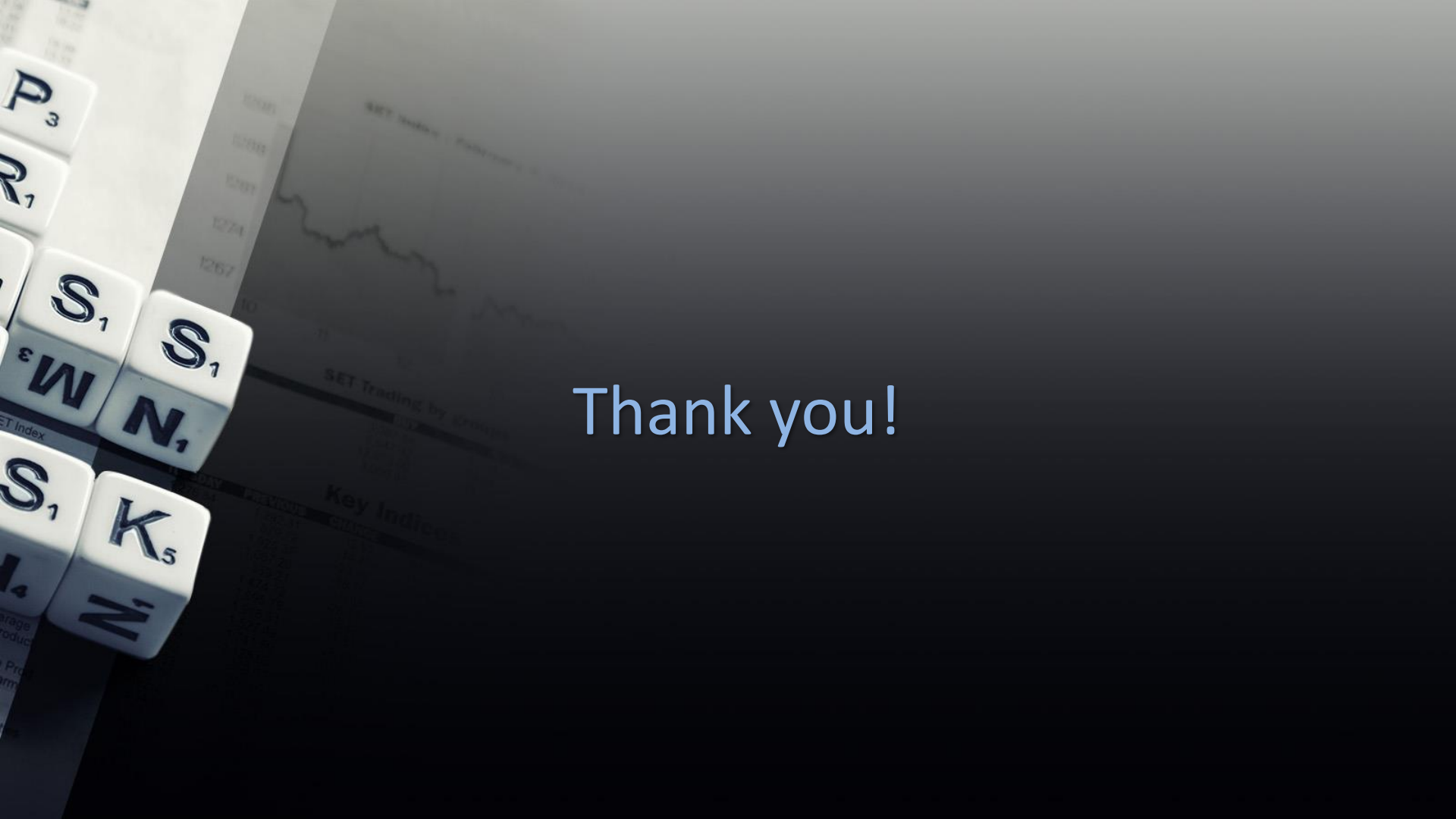


Conclusions

- The 3 classification models achieved high-level predictions (KNN 84.4%, Decision Tree 92.4%, Random Forest 95.5%) and a large amount of correct classifications.
- Especially, Random Forest classifier had the best overall performance in this analysis.
- Machine learning techniques can be developed and applied to banking and financial to improve their operations and drive business decisions.

Conclusions

- Although Machine Learning is considered as a useful tool for credit risk analysis and default prediction, there are several limitations connected to this type of analysis with the most important being the data quality and predictive strength.
- As economic conditions are continuously and rapidly changing, whereas new customers, new products and new trends are being introduced, new features, variables and correlations are needed to be considered.


Thank you!