**SUMMER SCHOOL ON RESEARCH METHODS 2020**

**STATUS OF A BANK LOAN**

**КАЗУС**

****

**ИЗГОТВИЛИ: МИХАЕЛА АНГЕЛОВА, ЯНА НЕЙКОВА,**

**ДИМИТЪР ТРИЧКОВ, АЛЕКСАНДЪР БАЛАБАНОВ**

**2020 г.**

**СЪДЪРЖАНИЕ:**

[**1.** **КРАТКО РЕЗЮМЕ НА ОСНОВИТЕ РЕЗУЛТАТИ** 3](#_Toc46624627)

[**1.1.КОНТЕКСТ** 3](#_Toc46624628)

[**1.2.ЦЕЛ** 4](#_Toc46624629)

[**1.3.ПОСТИГНАТИ РЕЗУЛТАТИ** 4](#_Toc46624630)

[**2.** **ПОДРОБНО ОПИСАНИЕ НА АНАЛИТИЧНИЯ ПРОЦЕС** 4](#_Toc46624631)

[**2.1.ДАННИ** 4](#_Toc46624632)

[**2.2.ОБРАБОТКА НА ДАННИТЕ** 5](#_Toc46624633)

[**2.2.1.ПРИЛОЖЕНИ ТРАНСФОРМАЦИИ** 5](#_Toc46624634)

[**2.2.2.ГЕНЕРИРАНИ СИНТЕТИЧНИ ХАРАКТЕРИСТИКИ** 5](#_Toc46624635)

[**2.3.ИЗБОР НА СЕГМЕНТИРАЩИ ХАРАКТЕРИСТИКИ** 5](#_Toc46624636)

[**2.4.ИЗПОЛЗВАНИ АНАЛИТИЧНИ ТЕХНИКИ** 6](#_Toc46624637)

[**2.5.ПЪРВОНАЧАЛНИ АНАЛИЗИ** 8](#_Toc46624638)

[**2.6.ИНТЕРПРЕТАЦИЯ НА ПОЛУЧЕНИТЕ РЕЗУЛТАТИ ОТ ВАЛИДИРАНЕ НА ПРОГНОЗНИТЕ МОДЕЛИ** 10](#_Toc46624639)

[**2.7.ОГРАНИЧЕНИЯ НА ПРОВЕДЕНОТО ИЗСЛЕДВАНЕ** 13](#_Toc46624640)

[**3.** **ПРИЛОЖЕНИЕ НА МОДЕЛА** 14](#_Toc46624641)

# **КРАТКО РЕЗЮМЕ НА ОСНОВИТЕ РЕЗУЛТАТИ**

# **КОНТЕКСТ**

Бизнесът, в лицето на банката, е изправен пред предизвикателството да реши проблема с големия брой клиенти, които не изплащат своите кредити. За да вземе адекватно решение как да предотврати задълбочаването на този проблем, бизнесът следва да разбере кои са основните фактори и предпоставки за появилата се ситуация.

Необходим е задълбочен анализ на клиентите по отношение на различни аспекти и идентифициране на слабите места и възможностите за развитие на банката с цел решаване на проблема.

В следствие на проведените анализи са формулирани следните групи, склонни към непогасяване на своите вноски:

* клиенти, които имат просрочени вноски през последната година
* клиенти с повече от три кредитни проблема
* клиенти с много висок кредитен рейтинг
* клиенти с дългосрочни заеми
* клиенти, които имат собствена къща, рента или начална ипотека
* клиенти, които са теглили бизнес заем или кредит за собствен бизнес или преместване

В резултат на проведеното изследване са идентифицирани следните фактори и предпоставки за не/изплащане на кредитите



# **ЦЕЛ**

Целта на настоящата разработка е да се създаде прогнозен модел за предсказване броя на клиентите на банката, които няма да изплатат своите кртедити.

В проекта за зададени следните подцели:

* Анализ на факторите за неизплащане на кредита с цел минимизиране на бъдещи рискови клиенти
* Създаване на прогнозен модел с добавена стойност
* Интегриране на решението в стратегията на банката

# **ПОСТИГНАТИ РЕЗУЛТАТИ**

В резултат на проведеното изследване са постигнати следните резултати:

* Обработени и изчистени данни
* Анализ на факторите и предпоставките, които оказват влияние върху изплащането на кредитите
* Формулирани синтетични характеристики и характеристики, отчитащи синергичен ефект между няколко фактора
* Избрани най-добри предсказващи характеристики
* Проверка на хипотези
* Проверка на статистическата значимост на избраните характеристики
* Приложени различни аналитични техники за създаване на прогноза
* Формулирани и оценени модели в следствие прилагането на различни аналитични техники
* Сравнителен анализ на създадените модели
* Избран окончателен прогнозен модел за броя отказващи се клиенти

# **ПОДРОБНО ОПИСАНИЕ НА АНАЛИТИЧНИЯ ПРОЦЕС**

# **ДАННИ**

Изходната информация за разработване на решение по казуса включва описание на задачата и една таблица с данни с характеристиките на клиeнтите.

Данните включват една целева характеристика /Loan.Status/ и 16 предсказващи характеристики.

# **ОБРАБОТКА НА ДАННИТЕ**

# **ПРИЛОЖЕНИ ТРАНСФОРМАЦИИ**

**Приложените обработки, свързани с трансформации на данните, са следните:**

* Коригиране формата на данните
* Изчистване на дубликати и запълване на липсващи стойности
* Изтриване на редове с голям брой лисващи стойности
* Изтриване на нетипични наблюдения /outlyers/
* Анализ на честотни разпределения
* Анализ на описателни статистики
* Боксдиаграми за визуализиране на наблюденията
* Създаване на факторни променливи – биниране

# **ГЕНЕРИРАНИ СИНТЕТИЧНИ ХАРАКТЕРИСТИКИ**

Преди да преминем към създаване на прогнозен модел, извършваме анализ на обясняващите характеристики. От проведените анализи, създаденото класификационно дръвче и формулираните изводи, създаваме няколко факторни променливи с по няколко нива в зависимост от описателните статистики и таблици с пропорции на платили/неизплатили своя кредит клиенти, които ни помагат да видим кои характеристики имат добра дискриминираща сила. По този начин бинираме голяма част от характеристиките, като формулираме синтетични променливи.

Последната променлива е изведена като комбинация от фактори, т.е. с тях се отчита синергичния ефект от съвкупната реализация на два фактора, а именно съотношението между месечен дълг и месечен доход, получен при разделяе на годишния доход на 12 месеца.

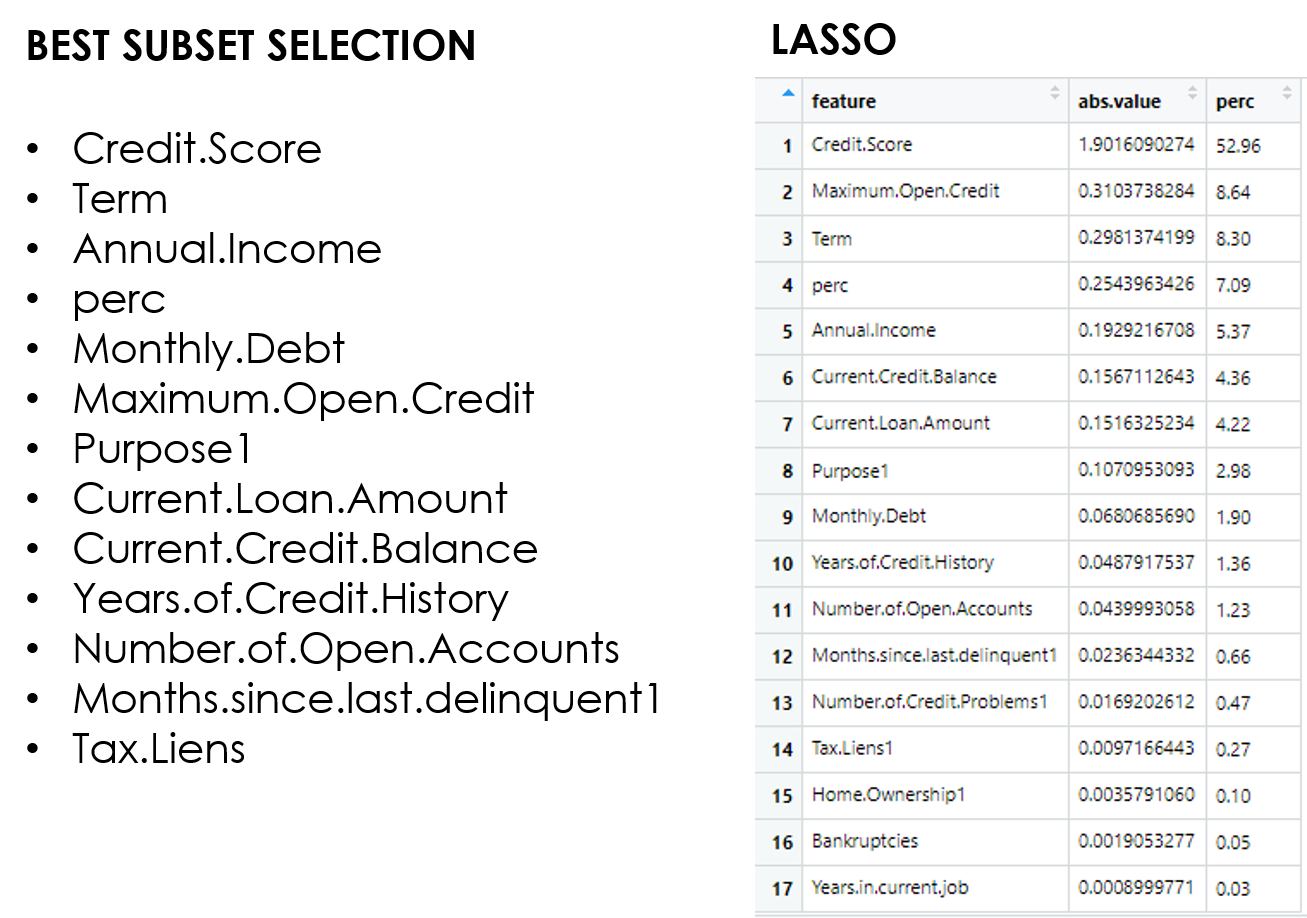
# **ИЗБОР НА СЕГМЕНТИРАЩИ ХАРАКТЕРИСТИКИ**

В рамките на тази задача използваме съкратени алгоритми за избор на най-добрите предсказващи характеристики. Посредством този метод получаваме като резултат, че оптималния брой предсказващи характеристики е 12. След прилагане на тези съкратени алгоритми проверяваме дали избраните характеристики са статистически значими, след което ги включваме в регресионно уравнение.

Тестваме и още една техника за избор на най-важни предсказващи характеристики посредством намаляване на размерността на множеството /Lasso/.

След прилагане на тази техника отново проверяваме дали избраните характеристики са статистически значими. Правим сравнение между двата метода за избор на предсказващи характеристики.

Виждаме, че имаме съвпадние и в двата алгоритъма.



Посочените методи ни показват кои са най-важните предсказващи характеристики за създаване на прогнозния модел.

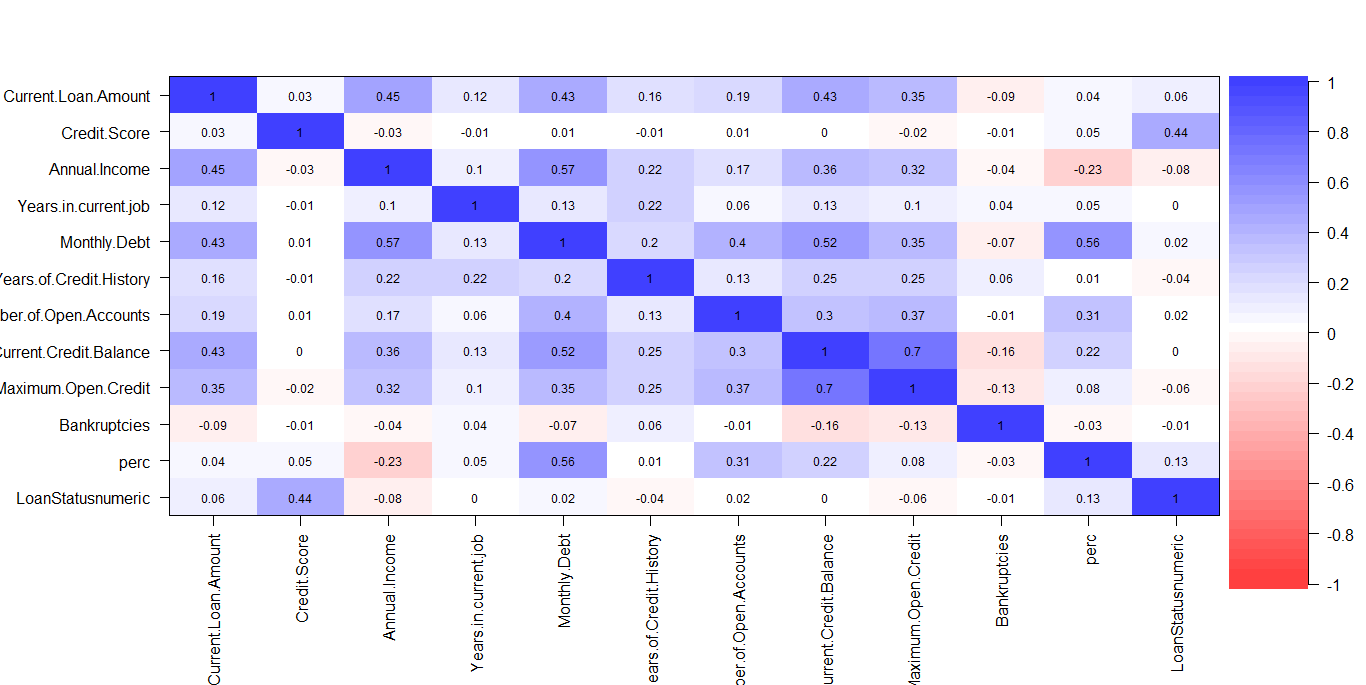
# **ИЗПОЛЗВАНИ АНАЛИТИЧНИ ТЕХНИКИ**

При разработване на настоящият проект са използвани следните аналитични техники:

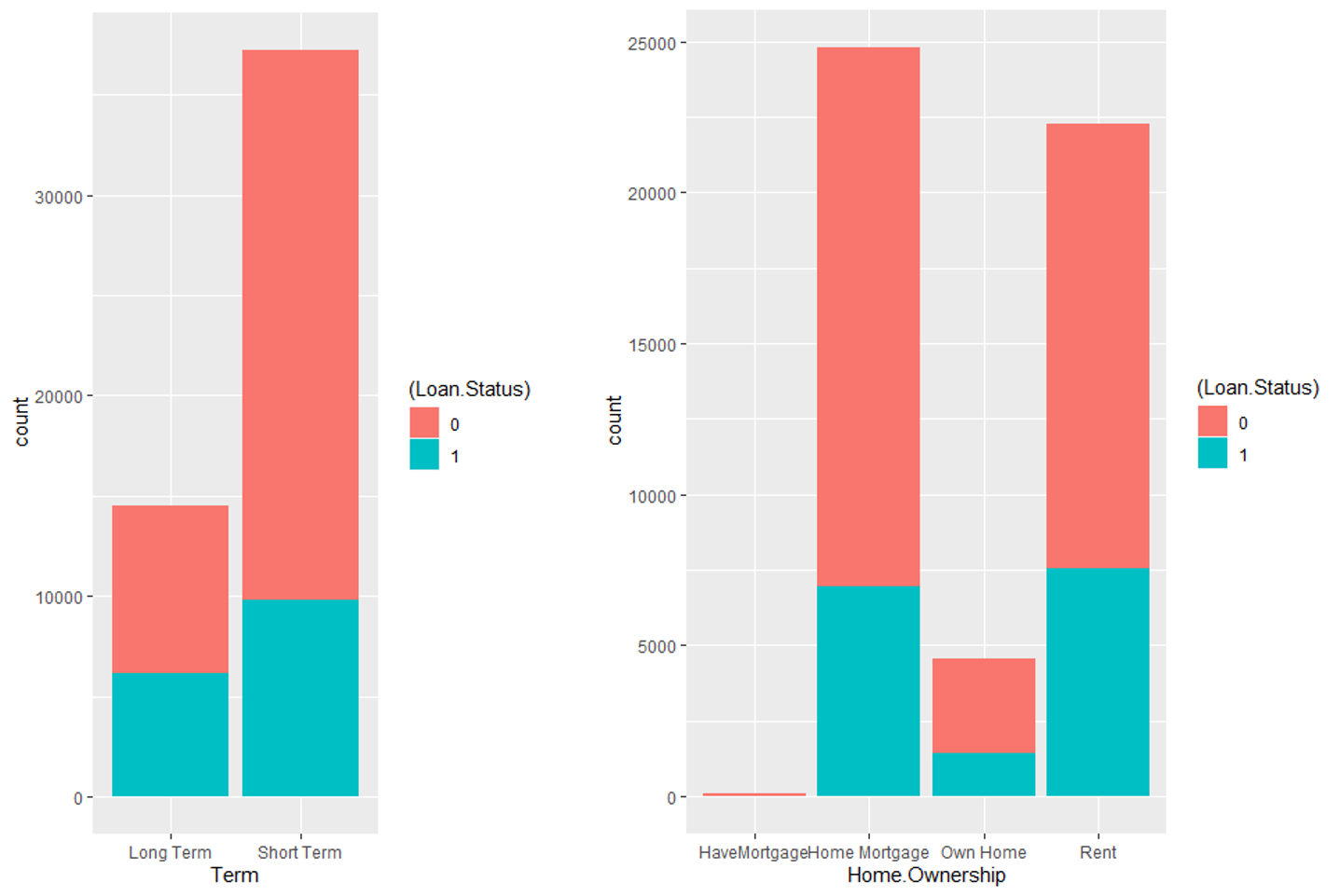
* **Боксграфики и таблици с пропорции** на неплатилите по различни характеристики, с цел идентифициране на дискриминиращи характеристики, визуализиране на целевата променлива и предсказващите характеристики, за изследване на нетипични наблюдения.
* **Класификационно дръвче /decision tree/** за идентифициране на най-важните предсказващи характерстики
* **Създаване на синтетични характеристики** **/Feature engineering/** и характерситики, които отчитат синергичен ефект, които достатъчно добре да обясняват модела
* **Алгоритъм за избор на най-добрите предсказващи характеристики** **/Best subset selection/**, който избира вместо нас кои са най-важните предсказващи характеристики и тяхното степенуване по важност
* **Техника за избор на най-добрите предсказващи характеристики /Lasso/**, която избира вместо нас кои са най-важните предсказващи характеристики като има дава процентна стойност и компресира в 0 незначителните такива
* **Логистична регресия** **/Logistic Regression/** – за създаване на прогнозен модел за класифициране на наблюденията. Тази техника е много полезна за проверка на статистическа значимост на характеристиките и интерпретация на зависимостите между целевата характеристика и предсказващите характеристики посредством интерпретация на β-коефициентите.
* **Линеен дискриминантен анализ** **/Linear discriminant analysis/** за създаване на прогнозен модел за класифициране на наблюденията и формулиране на дискриминиращи характеристики.
* **Квадратичен дискриминантен анализ** **/Quadratic Discriminant Analysis/** за създаване на прогнозен модел за класифициране на наблюденията и формулиране на дискриминиращи характеристики.
* **Техниката случайна гора /Random Forest/**, при която се генерират и оценяват едновременно множество независими едно от друго дръвчета дръвчета, с цел осредняване на резултатите и повишаване точността на крайния модел.
* **Визyализации на значимостта на характеристиките /Variable importance plots/** за идентифициране на предсказващите характеристики по степен на важност
* **Техниката Extreme Gradient Boosting /Xgboost/** за отглеждане на голям брой малки дръвчета, като всяко следващо дръвче взема резултатите от предходното, като бавно и постепенно подобрява прогнозната точност там където има проблем.
* **Сравнителен анализ** между различните техники за избор на предсказващи характеристики с цел постигане на максимално точен прогнозен модел
* **Матрици на объркването /Confusion matrix/** за получаване на информация за правилно и грешно прогнозираните стойности по класове. Тя показва как класификационния модел се „обърква“.
* **Оценка на грешката /TER/** на моделите в резултат на прилагането на различните техники с цел избор на най-добър прогнозен модел
* **Хистограми** за графично изобразяване на резултататите – визуализация на разпределението на различния % оценена вероятност клиентите да се откажат

# **ПЪРВОНАЧАЛНИ АНАЛИЗИ**

Корелационната матрица на числовите характеристики не показва високи корелации между целевата и предсказващите характериситки. Единствената по-висока корелация е с кредитния рейтинг.

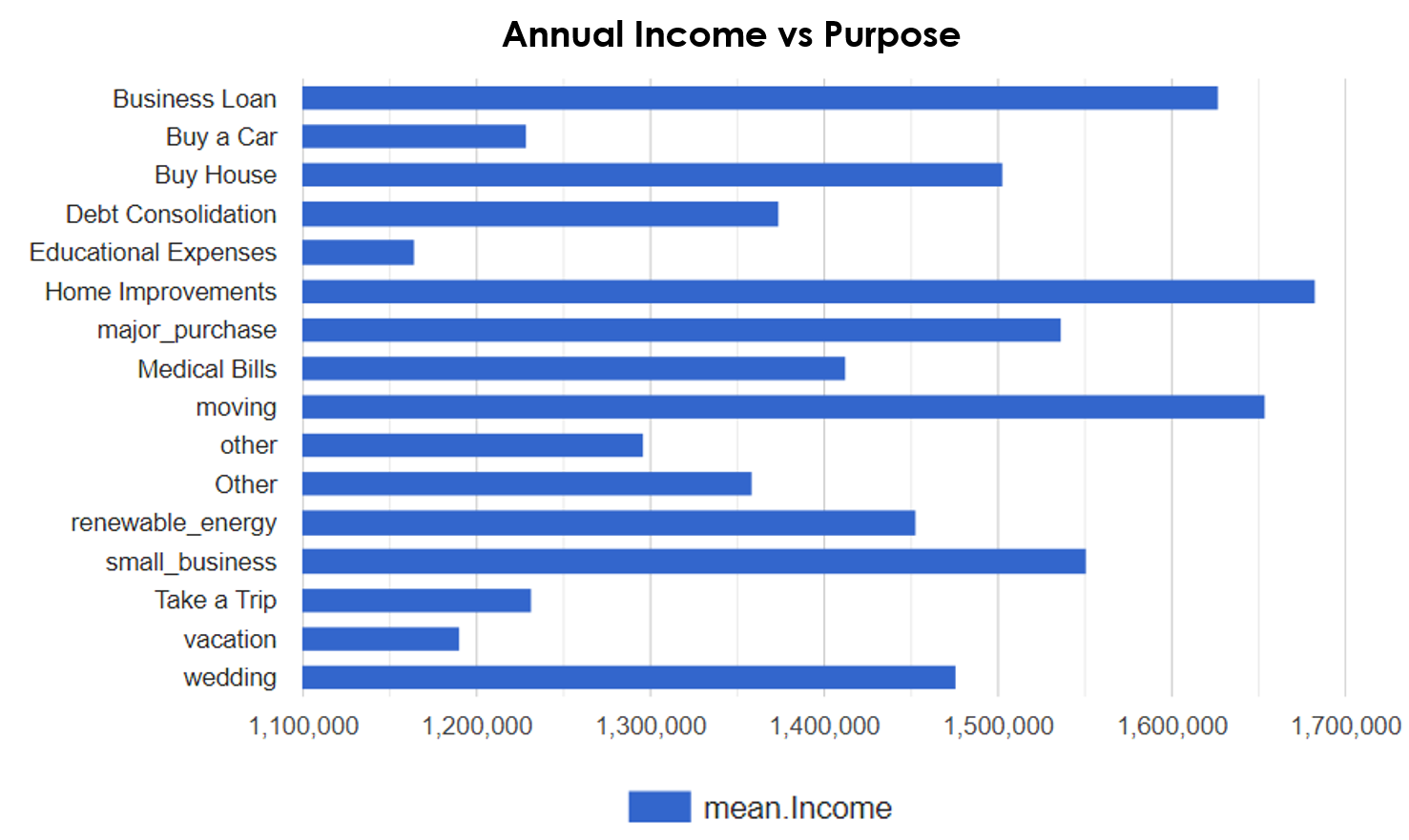


Посредством боксграфики, за всички качествени характеристики сме идентифицирали сегментиращата сила на двата класа.

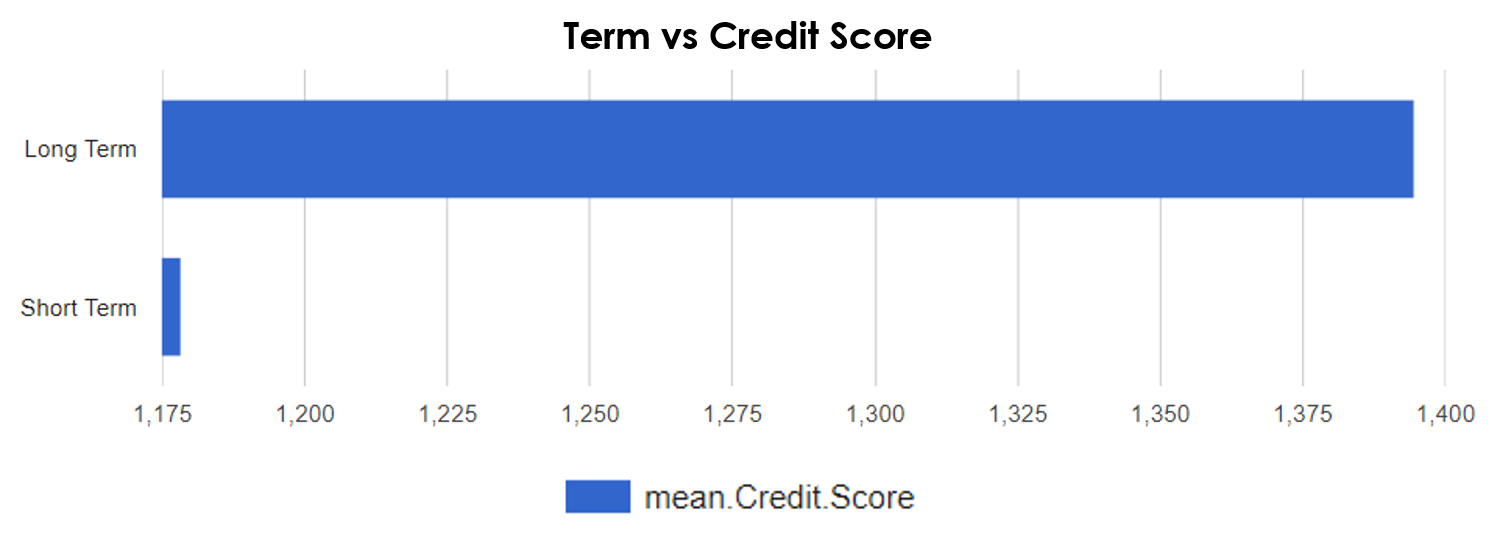


В послвдваствие сме създали факторни /дъми/ променливи с няколко нива, за да повишим обясняващата сила.

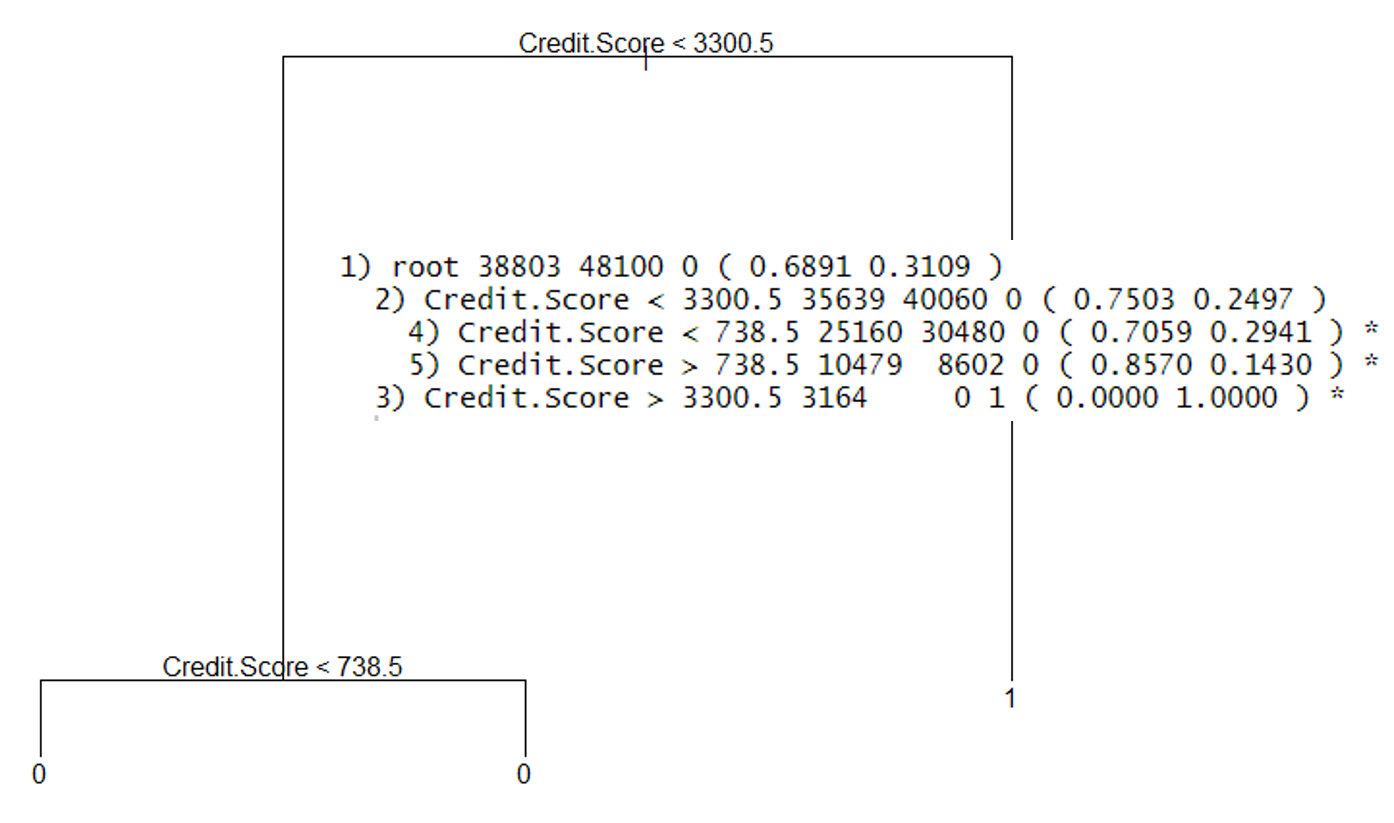
Анализът показва, че клиентите с различен годишен доход се групират около различни цели на заемите, като теси с по-висок доход теглят кредити предимно за развитие на бизнес идеи.



Краткосрочният заем е предпочитан от по-голямата част от кандидатите за кредит . Клиентите главно кандидатстват за заем за консолидация на дълга. Срокът на заема зависи от кредитния рейтинг на клиента



Дървото на решенията показва, че кредитният рейтинг е най-важната предсказваща характеристика.

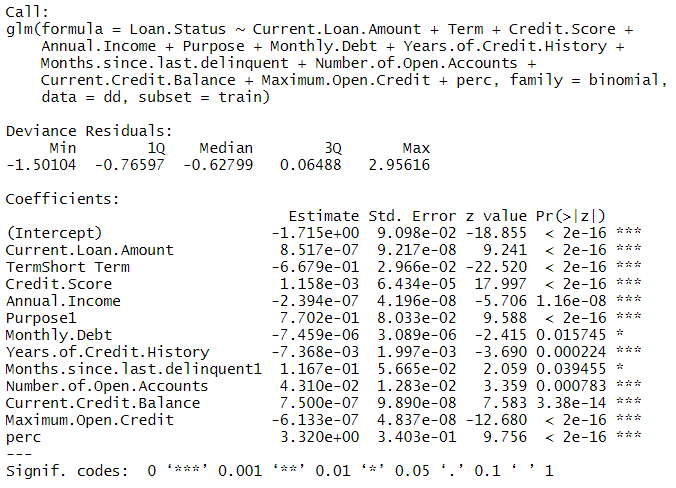


# **ИНТЕРПРЕТАЦИЯ НА ПОЛУЧЕНИТЕ РЕЗУЛТАТИ ОТ ВАЛИДИРАНЕ НА ПРОГНОЗНИТЕ МОДЕЛИ**

* **Интерпретация на получените резултати от прилагане на различните модели**

В следствие прилагането на различните прогнозни модели интерпретираме получените β-коефициенти и графики на значимостта на предсказващите характеристики.

С отрицателни тегла виждаме фактори благоприятстват за **намаляване** на вероятността, а с положителни – факторите, които допринасят за **нарастване** на вероятността клиентите да неизплатят кредита си.



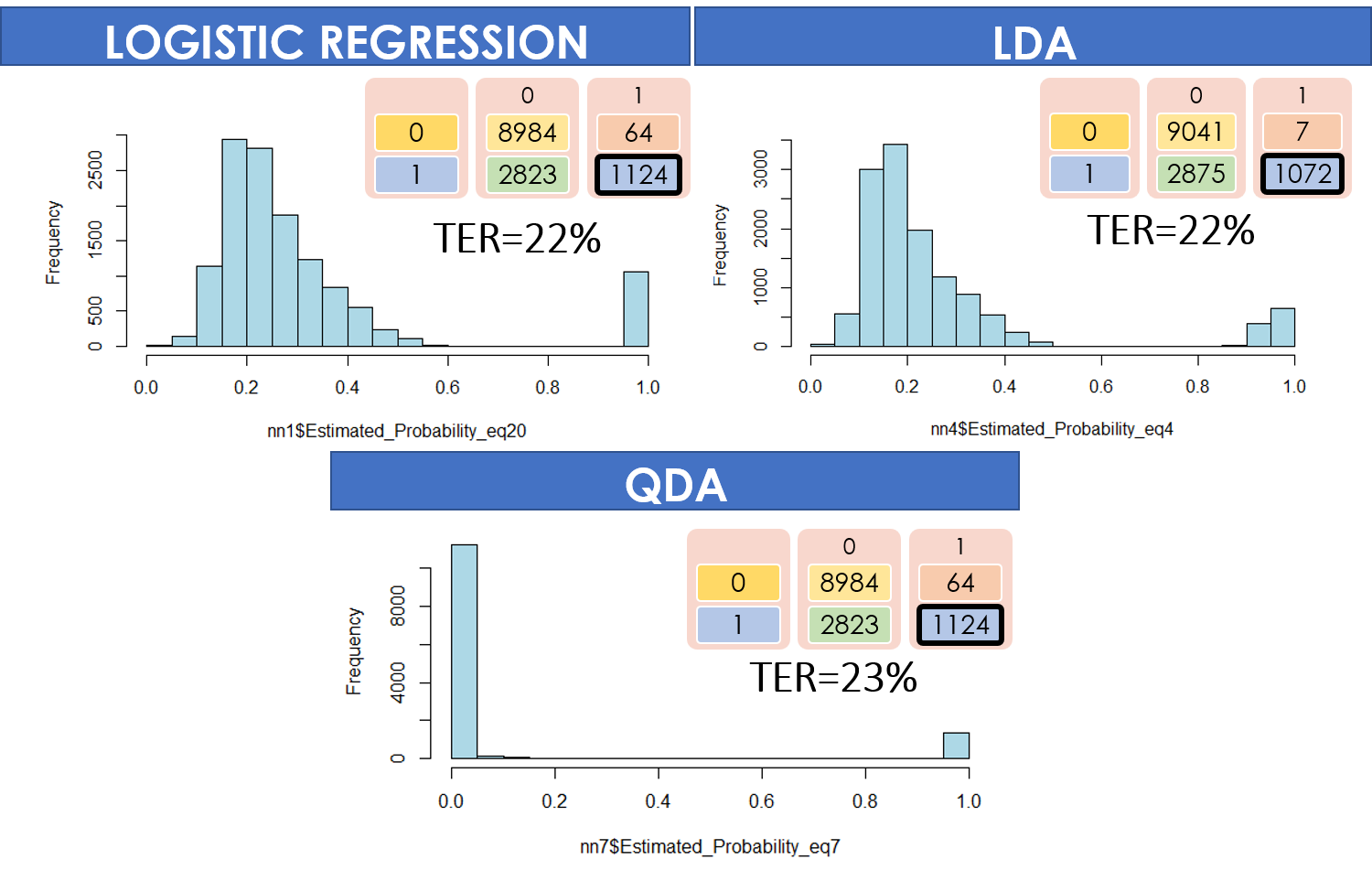
* **Интерпретация на получените резултати в следствие на проверка на грешките** **/TER/ на моделите, генерирани при прилагането на различните техники**

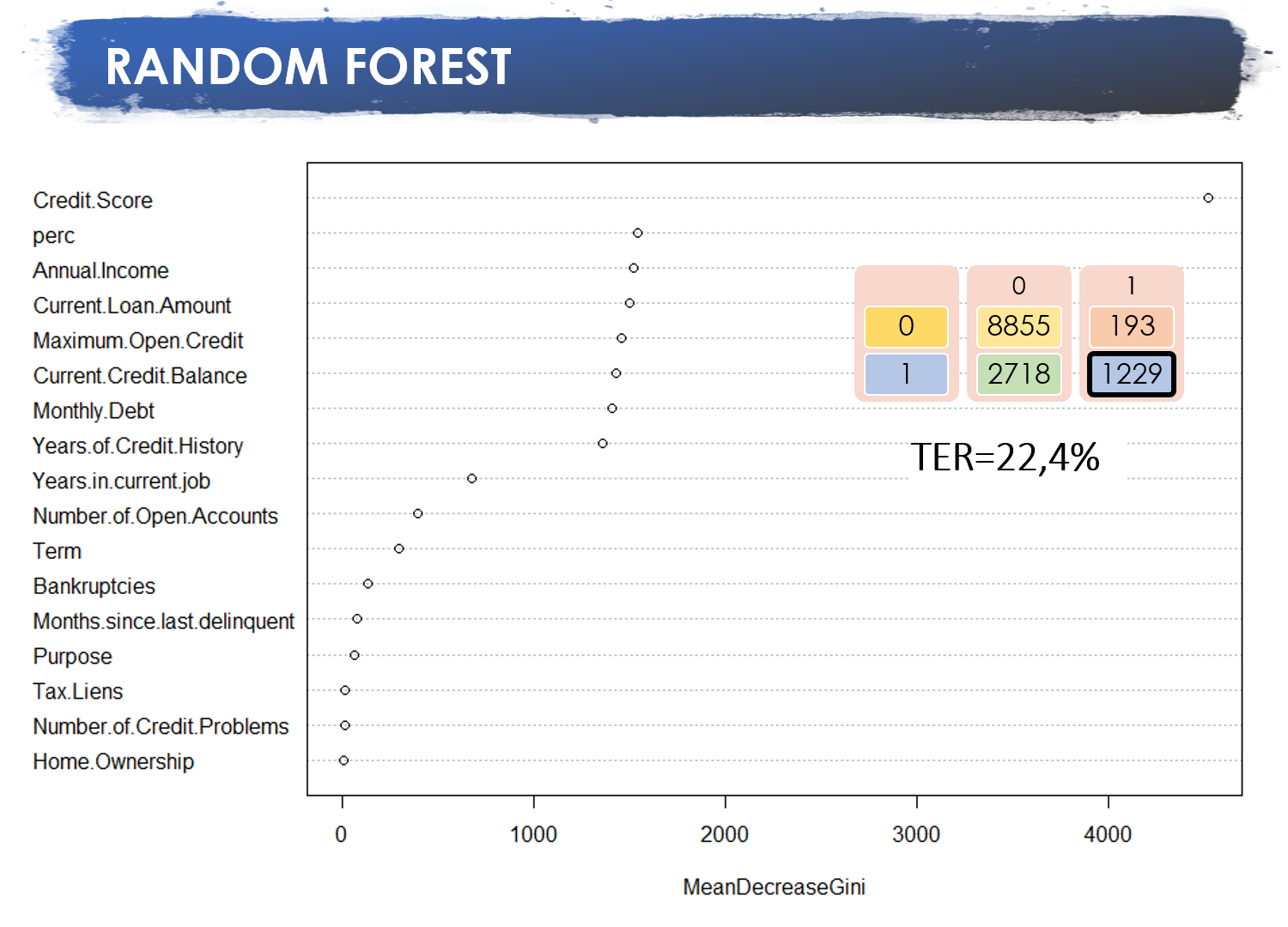
В резултат на валидирането на различните прогнозни модели за предсказване броя на клиентите, които няма да платат дълга си, сме извършили сравнителен анализ на прогнозната точност последством сравняване на грешките, които допускат моделите. По този начин избираме прогнозен модел с най-висока точност.

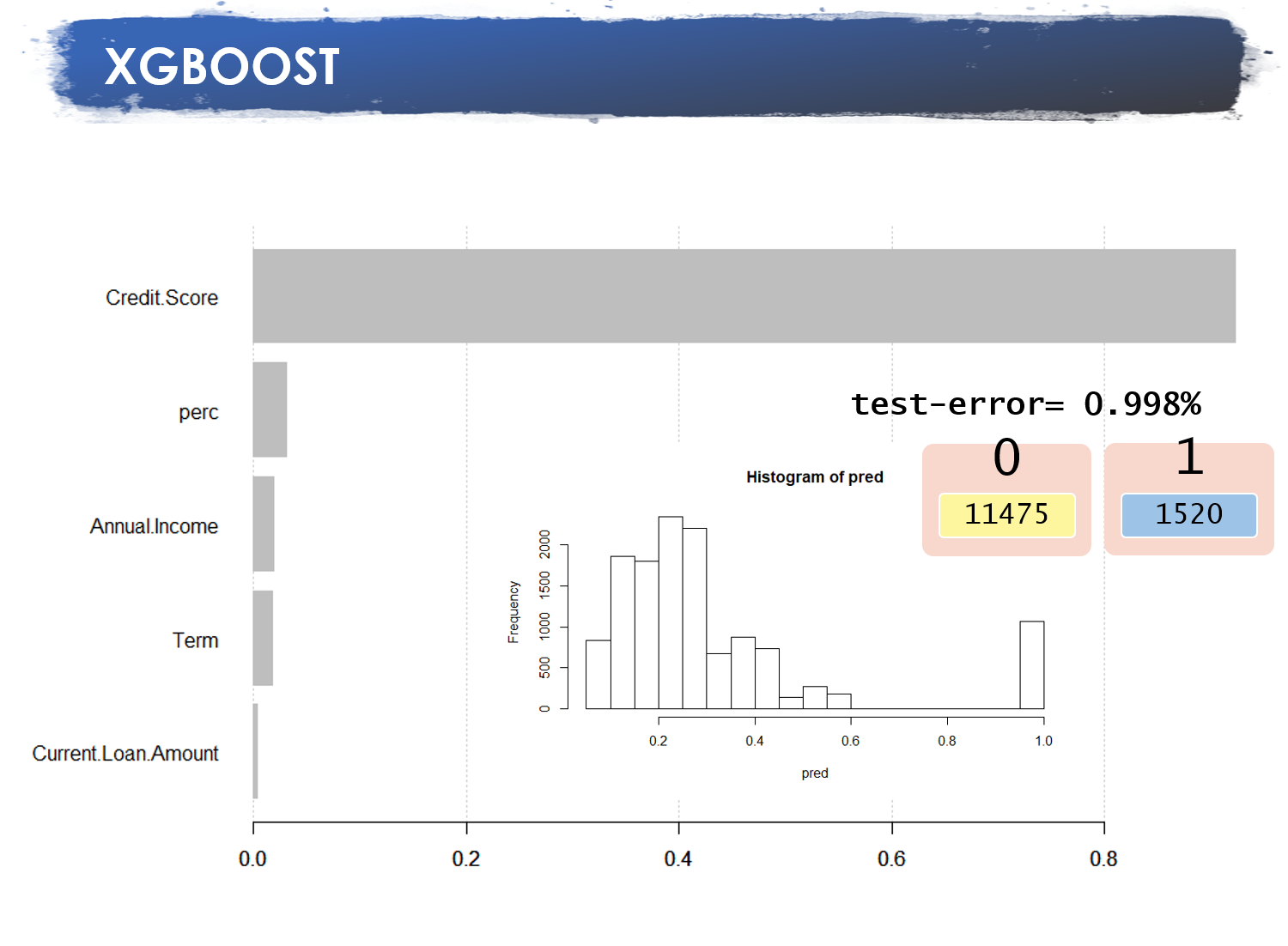
За целта отпечатаме матрицата на объркване confusion matrix за всеки от моделите и изчисляваме грешката.

Матрицата на обърквания/несъответствия /confusion matrix/ показва правилно и грешно прогнозираните стойности по класове и представя накратко цялостното изпълнение на класификационния алгоритъм. Тя показва как класификационния модел се „обърква“.

Резултати от тестваните модели:







Най-ниска е грешката при моделa с Xgboost – под 1% грешка. Този модел познава и най-много клиенти, които няма да изплатят кредита си.

Следователно моделът, оценен с Xgboost ще е и окончателният ни модел, който много точно разпознава добрите от лошите клиенти.

# **ОГРАНИЧЕНИЯ НА ПРОВЕДЕНОТО ИЗСЛЕДВАНЕ**

Настоящото изследване е проведено при следните ограничения:

* Липса на демографски данни за клиентите: възраст, пол, партньор
* Липса на баланс в извадката по отношение на някои предсказващи характеристики, което е предпоставка за недостатъчно точна оценка на дискриминиращата сила на тези характеристики
* Липса на баланс в извадката по отношение на целевата характеристика, клиентите които не са изплатили своя кредит са около 30%, т.е. около два пъти по-млако от редовно плащащите. Небалансираната извадка, може да е причина, моделът да работи и да разпознава по-добре наблюденията от класа, който доминира, и да не дава точна прогноза.

# **ПРИЛОЖЕНИЕ НА МОДЕЛА**

Създаденият прогнозен модел ще бъде от полза на банката по отношение на:

* класифициране и анализ на кандидатите за кредит
* помагане на банкерите да разберат поведението на клиентите
* идентифициране на добри и лоши /рискови/ клиенти
* даване на кредити в „сигурни ръце“