<<Mortgage Fairness>>
Un’analisi di equità nelle richieste di mutuo

Introduzione

L’obiettivo del nostro studio è quello di comprendere se negli anni 2015 e 2020 nello stato di New York vi siano state tendenze discriminatorie, per genere e/o etnia, riguardanti le richieste di mutuo.

Abbiamo scelto questo argomento perchè interessate all’argomento XAI(Explicable AI), che si occupa di spiegare in un modo comprensibile agli umani le decisioni e gli output di complessi sistemi di machine learning, chiamati modelli black box.
Gli algoritmi di XAI seguono tre principi chiave: transparency, interpretability and explainability. La differenza tra i tre concetti non è netta e definita, ma in linea teorica la transparency è definita come la capacità, da parte degli esperti di dominio, di comprendere appieno la teoria e la matematica sottostanti al modello, mentre la interpretability si riferisce alla possibilità di comprendere le ragioni per cui il modello black box effettua determinate scelte e dà come risultati determinati output. Infine la explainability viene definita come la capacità di comprendere appieno il modello nel suo insieme.

Alla luce di queste nozioni abbiamo scelto di analizzare due differenti dataset, uno riguardante il 2020 e l’altro riguardante il 2015, delle richieste di prestito per un mutuo nello stato di New York.

Metodologia

Le reti neurali sono dei sistemi di machine learning che usano un insieme di funzioni(percettroni), che date delle variabili di input generano un output, generalmente in forma binaria. La complessità delle rete neurale è solitamente data dal numero di strati e dal numero di neuroni in ogni singolo strato. Sono tra gli algoritmi di machine learning meno comprensibili agli umani, i cosiddetti modelli black box, perché è impossibile per una persona tenere conto di ciò che succede in ogni singolo strato. Sono però tra i migliori tools di machine learning per le loro buone, se non ottime, performance soprattutto con dataset di grandi dimensioni. Per queste ragioni abbiamo scelto come modello black box proprio una rete neurale.

Per interpretare la rete neurale abbiamo scelto due diversi modelli surrogati; un albero decisionale e SHAP.

Un modello surrogato è un modello che sostituisce e traduce il modello originale ed è in grado di renderlo interpretabile e di rappresentarlo al meglio. E’ utilizzato in contesti in cui è essenziale comprendere ed interpretare le decisioni del modello black box, perchè queste hanno particolari ripercussioni nella vita reale. Ad esempio, per poter capire al meglio le decisioni di un software di “self-driving” cars, nel caso in cui avvenga un incidente. In queste situazioni infatti, è di assoluta importanza capire cosa non abbia funzionato per poter migliorare il software e prevenire futuri incidenti. Oppure, come nel nostro caso in studio, bisogna comprendere se la richiesta di un eventuale mutuo viene rifiutata per ragioni

discriminatorie come la nazionalità, l'etnia, il sesso e l'eta. Oltre che immorali, queste scelte, in alcuni paesi, sono anche perseguitate dalla legge. Perciò bisogna poter comprendere il modello black box per assicurarsi di rispettare anche le leggi vigenti.
L'albero decisionale è un algoritmo non supervisionato usato sia per problemi di classificazione che per problemi di regressione. Può essere usato sia come modello a sé stante, sia come parte integrante del preprocessing per altri modelli. Nella nostra analisi abbiamo utilizzato l'albero con entrambe queste differenti finalità.
I maggiori vantaggi di questo algoritmo di machine learning sono che è veloce, ma soprattutto è facilmente interpretabile dato che fornisce delle regole di immediata comprensione. E' principalmente questo il motivo che ci ha portate a scegliere l'albero come modello surrogato anziché altri modelli con gli stessi vantaggi dell'albero, ma di più difficile interpretabilità, come i metodi di ensemble, dove la combinazione di più alberi porta ad una minore comprensione dell'algoritmo stesso.

SHAP (SHapley Additive exPlanations) è un metodo matematico usato per spiegare i risultati di un algoritmo di machine learning. Si basa sulla teoria dei giochi e cerca di calcolare il contributo di ogni singola variabile per l'output. Il "gioco" è quello di riprodurre l'output, mentre i "giocatori" sono le diverse features del modello. Ciò che SHAP cerca di fare è quantificare il contributo che ogni singolo "giocatore" apporta al "gioco", dove il gioco è un singolo output. Successivamente questi risultati vengono aggregati e come risultato i valori che SHAP fornisce indicano quanto è equamente distribuito l'output tra le diverse variabili del modello.
E' stato scelto perché al contrario di LIME(l'altro grande tool disponibile al momento per XAI) è un modello che fornisce delle spiegazioni globali, ossia che ci dice quali sono le variabili più importanti per il modello, che è ciò a cui siamo interessate.

Modelli locali e modelli globali

Esistono due diversi approcci per spiegare un modello black box; utilizzare un modello globale o un modello locale.
Nei modelli globali, come SHAP, ciò che il modello cerca di capire è come e quali sono i parametri che maggiormente influenzano il modello black box. Un modello surrogato di questo tipo quindi cerca di comprendere come il modello originale lavora nel complesso e quali fattori considera più importanti nella generazione dell'output. Per queste ragioni, i risultati di un modello globale non sono applicabili al singolo risultato del modello black box, ma sono generalizzabili. Uno degli svantaggi principali di questo tipo di modello è che se vi sono molte features un modello globale può avere delle basse performance.
I modelli locali invece, come LIME, hanno lo scopo di spiegare la singola decisione del modello, ossia perchè, per una determinata osservazione, è stato generato quel preciso output. I risultati dei modelli locali sono quindi specifici per una singola decisione e non generalizzabili a tutto il modello, come invece succede con i modelli globali. Il vantaggio di questi modelli è che possono portare ad una migliore comprensione di come e quanto determinate variabili influenzano un sottogruppo di osservazioni, informazione che utilizzando un modello andrebbe invece persa.
Dato lo scopo della nostra analisi, ossia comprendere se vi siano tendenze discriminatorie nelle richieste di mutuo, abbiamo scelto di utilizzare un modello surrogato globale, appunto perché desideriamo capire in base a quali fattori un prestito viene erogato o meno, quindi vogliamo avere una comprensione globale del modello black box.

Analisi

Ai fini della nostra analisi abbiamo considerato i dati dell'HMDA, *"Home Mortgage Disclosure Act"*, del 2015 e del 2020. L'HMDA è una legge vigente negli Stati Uniti d'America che ha lo scopo di raccogliere informazioni sulle differenti richieste di mutuo nel paese. Queste informazioni servono alle istituzioni per comprendere se vi sono alcune parti della nazione dove il mercato immobiliare è particolarmente in crisi, o se vi è bisogno di maggiori incentivi per comprare casa, e anche per individuare possibili tendenze discriminatorie nelle richieste di mutuo.
In particolare lo scopo della nostra ricerca è quello di concentrarci su quest'ultima analisi; vogliamo creare prima un modello black box e poi due differenti modelli surrogati per spiegare le decisioni del primo modello e vedere se queste sono biased, in particolare per sesso e/o etnia.

Come strumento di analisi dei dati abbiamo utilizzato "Google Colab", in particolare la libreria "Pandas" di Python.
Abbiamo cercato di mantenere il preprocessing e i diversi passaggi uguali per entrambi i dataset, anche se alcune volte sono state richieste alcune modifiche a seconda del dataset.

Analisi-2015

Per il 2015 avevamo a disposizione 78 variabili e 439654 osservazioni. Le osservazioni si riferiscono ad un campione estratto tra i richiedenti mutuo nel 2015 nello Stato di New York. I dati raccolti contengono informazioni socio-demografiche di coloro che chiedono il prestito, sulla casa in questione e sul proprietario della casa, sul tipo di istituzione a cui viene richiesto il mutuo, oltre che sul tipo di prestito richiesto.
La maggior parte delle variabili è presente in una doppia codifica "nominale"/ "ordinale" perciò abbiamo eliminato tutte le variabili con la codifica nominale, essendo la codifica ordinale migliore per le analisi statistiche nonché essenziale per alcuni particolari passaggi.

Abbiamo eliminato le variabili che presentavano l'80% o più di valori nulli e le variabili con valori costanti(come quelle che indicavano l'anno o il nome dello stato) e l'id. Inoltre alcune variabili presentavano moltissimi diversi valori perciò abbiamo effettuato un raggruppamento.

Il modello black box scelto per la nostra analisi è una rete neurale perciò gli step di preprocessing richiesti sono la gestione dei dati mancanti, la multicollinearità, la model selection, la near-zero variance e il centering e lo scaling.

Prima di occuparci della gestione dei dati mancanti abbiamo controllato la presenza di duplicati nel dataset e nel caso fossero presenti li abbiamo rimossi. Per la gestione dei dati mancanti abbiamo eliminato delle variabili che fornivano delle informazioni geografiche, questo perché ci sembrava che avesse poco senso fare una imputazione di dati mancanti su questo tipo di dato, sia perchè avevamo altre variabili che fornivano sempre delle indicazioni geografiche e che non presentavano dati mancanti.

Sulle restanti variabili che presentavano molti dati mancanti abbiamo effettuato una regressione logistica. In realtà la prima tecnica considerata per l'imputazione era un KNN(K-Nearest Neighbors) ma data la dimensione del dataset il KNN richiedeva troppe risorse computazionali.

Abbiamo inteso il target come una variabile binaria, coloro che ottengono il mutuo e coloro che non lo ottengono, ma la nostra variabile target presenta 7 diversi livelli, perciò abbiamo effettuato una codifica binaria.

Per la gestione della multicollinearità abbiamo deciso di utilizzare un test chi quadro, ma data la dimensione del dataset non è stato possibile utilizzarlo, perciò abbiamo effettuato una model selection tramite un albero decisionale per gestire il problema della multicollinearità. Come risultato l'albero considera due variabili come meno importanti, ma dato che non ne elimina nessuna decidiamo di procedere senza eliminare nessuna features.

Per la near zero variance abbiamo utilizzato come soglia il valore "0.1" ,vengono così eliminate due variabili.

Tutte le nostre variabili sono di tipo ordinale perciò non sono richiesti nè centering, nè scaling.

Rete neurale-2015

Abbiamo proceduto ad allenare la rete neurale sui dati di training e abbiamo visualizzato la matrice di confusione sui dati di test per vedere le performance del modello;

```
confusion_matrix_test_nn.head()
```

	Unnamed: 0	0.0	1.0	accuracy	macro avg	weighted avg
0	precision	0.648572	0.997364	0.807233	0.822968	0.873632
1	recall	0.996619	0.703114	0.807233	0.849867	0.807233
2	f1-score	0.785781	0.824780	0.807233	0.805280	0.810945
3	support	26327.000000	47887.000000	0.807233	74214.000000	74214.000000

Figura 1; Matrice di confusione sui dati di test, rete neurale, 2015

Come mostrato in Figura 1, la rete neurale ha un accuracy di 0.80 sui dati di test. Mentre la precision di 0.87 ci dice che, di tutti coloro che hanno davvero ottenuto il mutuo, noi abbiamo classificato correttamente l'87% di questo. In aggiunta, la recall pari a 0.80 ci dice invece che di tutti coloro che sono stati classificati come 1(ottenere il mutuo) il 80% l'ha davvero ottenuto. Il F1 score, che è una media armonica tra la precision e la recall, è pari a 0.81, indicando che il nostro modello ha generalmente delle buone performance sui dati di test.

Albero decisionale-2015

Dopodiché abbiamo allenato l'albero decisionale sugli stessi dati su cui abbiamo allenato la rete neurale. Come si evince dal grafico seguente, la matrice di confusione sui dati di test dell'albero ha delle performance sensibilmente migliori di quelle della rete neurale, ma molto simili; con un F1 score di 0.82, pari alla precision, alla recall e all'accurancy.

```
confusion_matrix_tree.head()
```

	Unnamed: 0	0.0	1.0	accuracy	macro avg	weighted avg
0	precision	0.742681	0.864839	0.821206	0.803760	0.821820
1	recall	0.753281	0.858129	0.821206	0.805705	0.821206
2	f1-score	0.747944	0.861471	0.821206	0.804707	0.821492
3	support	26135.000000	48079.000000	0.821206	74214.000000	74214.000000

Figura 2; Matrice di confusione sui dati di test, albero decisionale, 2015

Di seguito mostriamo un grafico che rappresenta le variabili più importanti per l'albero.

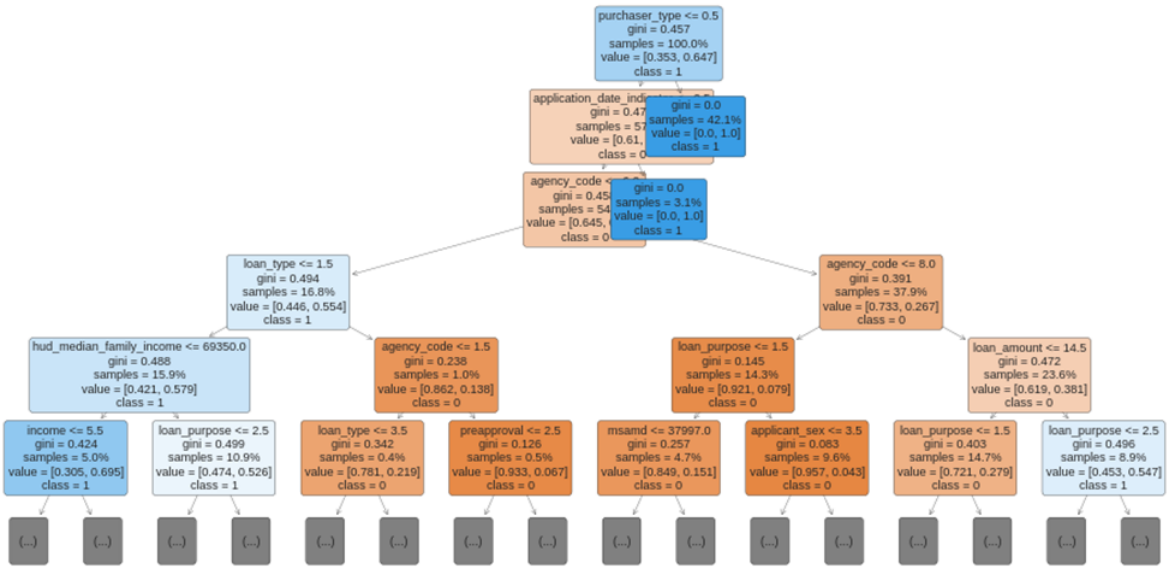

Figura 3; Variabili più importanti per l'albero decisionale, 2015

Le variabili più importanti che vengono individuate dall'albero sono *"purchaser_type"*, che era anche la variabile più importante secondo la model selection nel preprocessing. La variabile si riferisce, se e a quale istituzione il prestito è stato venduto. La soglia di "0.5", è importante perché coloro che hanno un valore minore sono coloro per cui il prestito non è stato venduto a nessun'altra istituzione, mentre altri valori indicano che il mutuo è stato venduto. Le banche, quando concedono un prestito, solitamente decidono a priori se venderlo ad una delle loro aziende sussidiare o altre istituzioni oppure no, e questa scelta sembra essere il fattore determinante nella concessione o meno del mutuo. Al secondo e terzo posto invece troviamo *"application_date_indicator"* e *"agency_code"*. La prima variabile indica quando è stata fatta la richiesta di mutuo e ha come soglia il 01.01.2015, ossia si

codificano diversamente le richieste di mutuo effettuate prima o dopo tale data. *"agency_code"* si riferisce al tipo di istituzione che invece sta facendo rapporto per l'HMDA.

Una possibile spiegazione per l'importanza della variabile *"application_date_indicator"* è la presenza di particolari incentivi statali per l'acquisto delle case, i quali potrebbero aver influenzato gli esiti delle richieste di mutuo effettuate in una certa data. La variabile *"agency_code"* invece, risulta importante non tanto perché è importante di per sé quale agenzia fa rapporto all'HMDA, ma l'agenzia che fa rapporto è quella a cui si fa la richiesta di mutuo, e questo quindi influenza molto l'esito della richiesta di mutuo, in particolare la differenza principale sembra essere tra agenzie federali e private. Come altre variabili importanti per l'albero troviamo *"loan_type"* con una soglia di "1.5", dove con valori minori di "1.5" vengono indicati i prestiti convenzionali, in contrapposizione con i prestiti garantiti dalle agenzie federali dello stato.

SHAP-2015

SHAP individua come importanti le medesime variabili trovate dall'albero nello stesso identico ordine. Al primo posto troviamo *"purchaser_type"* seguita da *"application_date_indicator"* e *"agency_code"*, come mostrato nel grafico seguente;

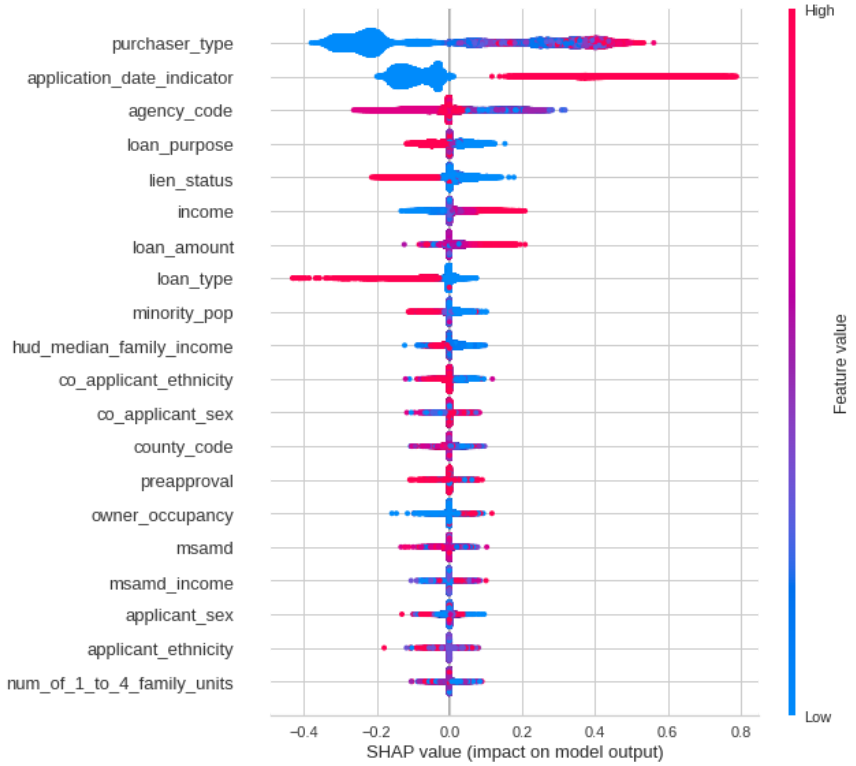

Figura 4; Variabili più importanti per SHAP, 2015

In particolare il modello sembra influenzato da valori bassi di *"purchaser_type"*, che indicano prestiti non venduti o venduti ad aziende finanziate dallo stato. Inoltre valori alti di *"application_date_indicator"*, i quali indicano le domande effettuate dopo il 01.01.2015 sembrano influenzare il modello. Infine si nota, come il modello è particolarmente sensibile a

bassi valori di *“loan_type”*, che identificano le richieste di mutuo effettuate per comprare la prima casa.

Di fronte a questi risultati sembra che per l’anno 2015 le richieste di mutuo nello stato di New York non siano state influenzate dal sesso o dall’etnia. Le variabili più importanti individuate sia da SHAP che dall’albero tendono ad essere fuori dal controllo del singolo richiedente, ma sono più legate a fattori esterni, come le decisioni della banca di vendere il prestito o la data entro la quale si fa la richiesta. Ora procediamo con le analisi per l’anno 2020.

Analisi-2020

Gli step eseguiti per il 2020 sono pressoché identici a quelli eseguiti per il 2015. Una differenza nei due dataset è la loro numerosità, per il 2020 avevamo a disposizione 720.000 osservazioni, quasi il doppio a quelle del 2015.
Si sono eliminati i dati con l’80% o più di dati mancanti, si sono raggruppate le variabili con troppi valori. Abbiamo eliminato i duplicati e usato una regressione lineare per l’imputazione di dati mancanti. La model selection con un albero decisionale ha individuato come variabile più importante sempre *“purchaser_type”* seguita però da *“denial_reason-1”,* ossia il motivo principale per cui viene negato il mutuo e *“income_class”* che indica in che fascia appartiene il reddito della persona che fa domanda di mutuo. Le variabili meno importanti sembrano essere quelle legate all’etnia. Per la near zero variance si sceglie sempre la soglia di “0.1,” e si eliminano cinque variabili.
Successivamente abbiamo allenato la rete neurale e visualizzato la matrice di confusione sui dati di test.

Rete neurale-2020

Le performance della rete neurale per il 2020 sono meno buone di quelle del 2015. Il preprocessing adottato è stato lo stesso, con le opportune modifiche per gestire il dataset. Abbiamo effettuato diversi raggruppamenti e provato diverse tecniche per la gestione dei dati mancanti, ma i risultati erano sempre simili. Perciò abbiamo tenuto questi risultati, anche se non ottimali. Il F1 score è pari a 0.50, mentre la metrica migliore risulta essere la recall con un valore di 0.64, pari all' accuracy.

```
confusion_matrix_test_nn.head()
```

	Unnamed: 0	0	1	accuracy	macro avg	weighted avg
0	precision	0.0	0.642704	0.642704	0.321352	0.413068
1	recall	0.0	1.000000	0.642704	0.500000	0.642704
2	f1-score	0.0	0.782495	0.642704	0.391247	0.502912
3	support	46018.0	82777.000000	0.642704	128795.000000	128795.000000

Figura 5; Matrice di confusione sui dati di test, rete neurale, 2020

Albero decisionale-2020

La matrice di confusione dell'albero sui dati di test è molto simile a quella ottenuta nel 2015 con un F1 score pari a 0.81, come la recall e la precision, come mostrato nell'immagine seguente, mentre l'accuracy è pari a 0.86;

```
] confusion_matrix_tree.head()
```

	Unnamed: 0	0	1	accuracy	macro avg	weighted avg
0	precision	0.730978	0.860936	0.812904	0.795957	0.814433
1	recall	0.755007	0.845166	0.812904	0.800086	0.812904
2	f1-score	0.742798	0.852978	0.812904	0.797888	0.813552
3	support	46087.000000	82708.000000	0.812904	128795.000000	128795.000000

Figura 6; Matrice di confusione sui dati di test, albero decisionale, 2020

L'albero decisionale individua come variabili più importanti per il modello *"purchaser_type"*, *"denial_reason-1"* e il *"applicant_sex"*, come mostrato dal seguente grafico;

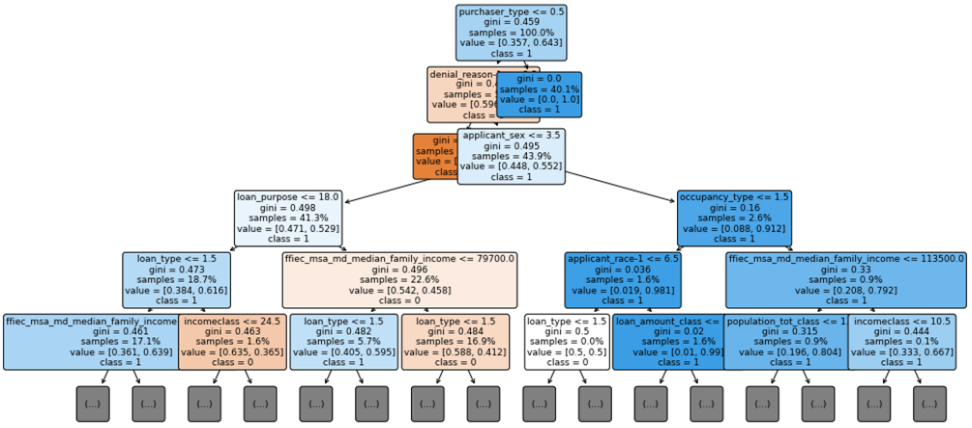

Figura 7; Variabili più importanti per l'albero decisionale, 2020

La prima variabile per importanza è la stessa identificata dall'analisi precedente, mentre *"denial_reason-1*" si riferisce al motivo principale per cui il mutuo non è stato concesso. *"applicant_sex"* si riferisce al sesso di colui/lei che fa richiesta di mutuo. Da questi primi risultati i dati del 2020 sembrerebbero biased per il sesso, ma l'albero sceglie come soglia i valori minori di "3.5", dove 1- maschio 2- femmina e i valori successivi significa che il dato non era disponibile o non era fornito. Questo significa che di per sè non c'era una discriminazione verso gli uomini o le donne, ma che fornire l'informazione sul proprio sesso, indipendentemente da quale esso sia, influisce sulla decisione finale.

SHAP -2020

SHAP individua come variabili più importanti sempre *"purchaser_type"* e *"denial_reason-1",* ma al terzo posto al contrario dell'albero, vi è la variabile *"loan_purpose"*, che indica perché si sta facendo la richiesta di mutuo; se per comprare un nuovo immobile, rinnovare la casa, rifinanziare un mutuo già esistente, ect. In particolare il modello sembra più influenzato dai valori bassi della features, che indicano prestiti richiesti per comprare la prima casa o rinnovarla, piuttosto che per rifinanziare un mutuo già esistente, perciò per quest'ultimo motivo si tende di meno a fornire dei prestiti. Successivamente come variabili importanti per il modello troviamo *"loan_amout_class"* che indica l'ammontare del prestito richiesto, per somme più basse si tende a concedere un prestito più facilmente.

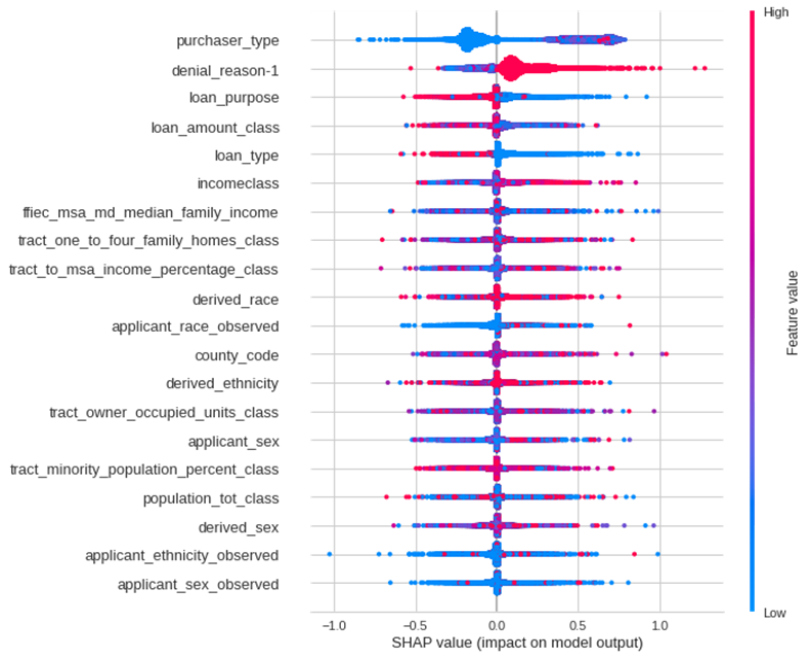

Figura 8; Variabili più importanti per SHAP, 2020

Conclusioni

Bisogna porre attenzione ai confronti dei risultati tra i due dataset perchè potrebbero essere influenzati dalla diversa numerosità dei due dataset. Ma alla luce dei risultati ottenuti possiamo supporre che non sembra esserci un bias per genere e/o etnia nelle richieste di mutuo nello stato di New York negli anni 2015 e 2020. Nei due differenti anni però si nota come la variabile più importante sia sempre *“purchaser_type”,* mentre nel 2015 si poneva più importanza alla singola agenzia a cui si faceva richiesta, mentre nel 2020 risultano essere più importanti il motivo per cui si chiede il prestito e l’ammontare dello stesso.

Bibliografia

- https://www.ffiec.gov/hmda/glossary.htm
- https://www.consumerfinance.gov/data-research/hmda/
- https://www.kaggle.com/
- [2011.07876] A Survey on the Explainability of Supervised Machine Learning (arxiv.org)